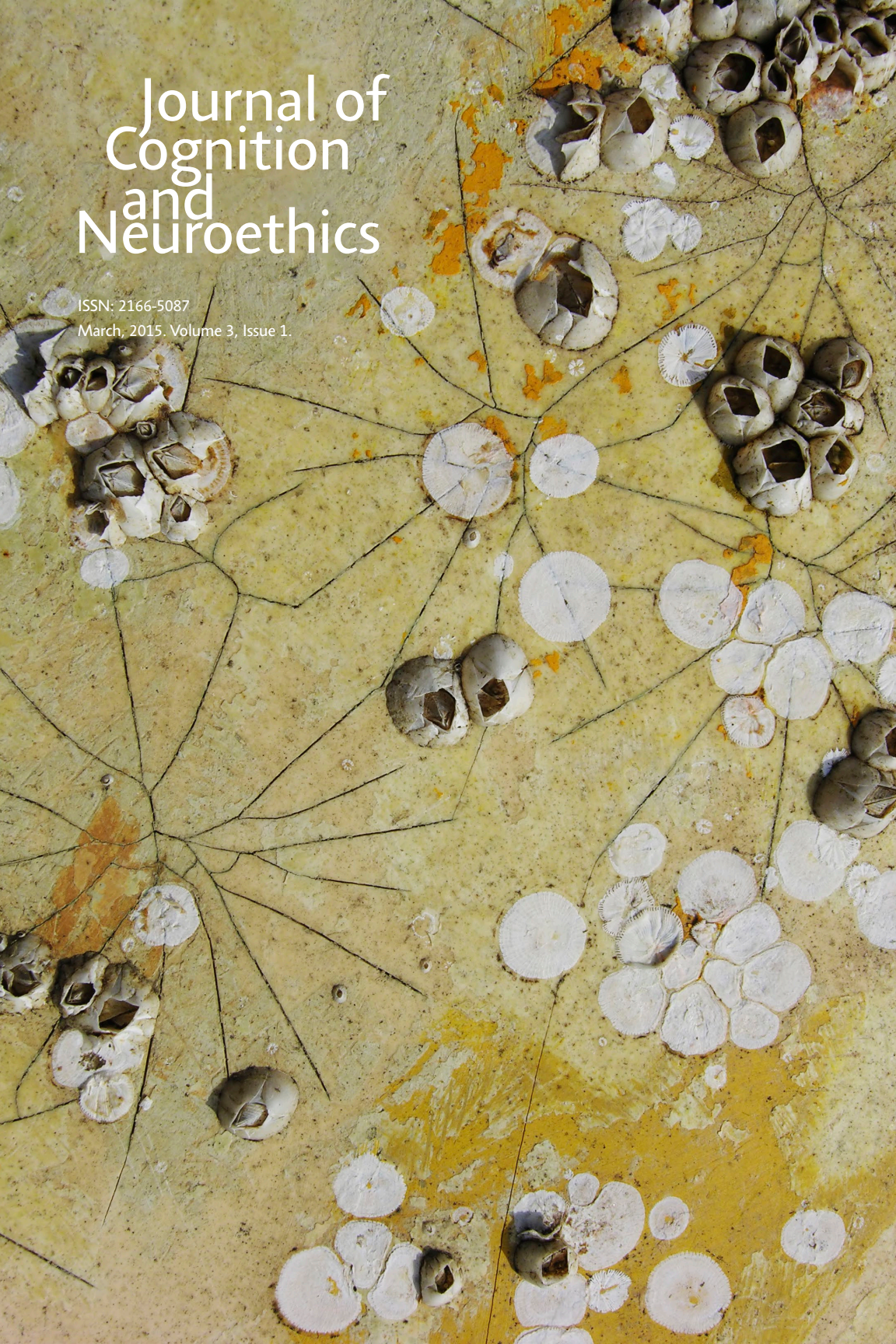
Journal of Cognition and Neuroethics

ISSN: 2166-5087
March, 2015. Volume 3, Issue 1.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 3, Issue 1 was digitally published in March of 2015 from Flint, Michigan, under ISSN 2166-5087.

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

Center for Cognition and Neuroethics
University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Journal of Cognition and Neuroethics

Is Islam Committed to Dualism in the Context of the Problem of Free Will?

Macksood Aftab
Harvard Extension

Biography
Dr. Macksood Aftab is a neuroradiologist, and clinical assistant professor at both Michigan State University and Central Michigan University. He holds a Master degree in History of Science, and is an editor for the *Journal of Islamic Philosophy*. The author can be reached at: mackaftab@post.harvard.edu.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Aftab, Macksood. 2015. "Is Islam Committed to Dualism in the Context of the Problem of Free Will?" *Journal of Cognition and Neuroethics* 3 (1): 1–12.

Is Islam Committed to Dualism in the Context of the Problem of Free Will?

Macksood Aftab

Abstract

The concept of *free will* became an early source of debate amongst theologians and philosophers within Islamic intellectual history. The influential theologian Ghazali refuted neoplatonic philosophers who laid claim to a deterministic view of human nature by pointing to the problem of causation. Ashari theologians refuted those who claimed humans create their own actions by arguing that such a statement would undermine God's omnipotence. The core issue in these Islamic debates was not so much *human* free will, but rather concern over the preservation of *divine* free will. A largely dualistic picture emerged regarding the nature of human beings from these early debates. Human free will did of course play an important role in law, morality and theology and was accommodated with the development of the doctrine of *kasb* (acquisition). It is not until recently, with the advances in neurophilosophy, psychology, neuroscience and neuroimaging that the Islamic theological understanding has required a revisiting. Recent thinkers such as Iqbal have critiqued Ghazali for not being able to break from dualism, and articulate a more wholesome view of man in which *human* free will takes center stage. I will argue that Iqbal's conception of free will within the Islamic context is not much different from that articulated by Daniel Dennett, even if the implications drawn by the two men are vastly different. The rich intellectual history of Islam is relevant to many of the contemporary debates on free will and these intersections will be discussed.

Keywords

Dualism, Free Will, Islam, Ghazali, Ibn Sina, Iqbal, Dennett, Soul

Soul & Free Will

The classic problem of free will has often assumed a worldview based on a dualistic perspective of the universe. This dualism which was once used to explain free will has now become an obstacle in the same discussion. According to this view, two distinct substances exist: physical and metaphysical. The body (the physical component) is acted upon by the soul (the non-physical component). The locus of the human capacity for free will is thought to be located in the soul (the non-physical component). A version of this type of dualism is also seen within Islamic intellectual history.

This particular view has come under sharp criticism by modern day scientists and philosophers. They argue that granting a metaphysical status to the soul exempts it from

being the subject of any meaningful scientific inquiry. If this is the case, then how can we be sure that it exists at all? Furthermore, if it exists, they ask, how can a metaphysical entity interact with a physical entity?

Daniel Dennett, the famous philosopher from Tufts University, summarizes the problem when he writes,

> One widespread tradition has it that we human beings are responsible agents, captains of our fate, because what we really are souls, immaterial and immortal clumps of Godstuff that inhabit and control our material bodies rather like spectral puppeteers. It is our souls, that are the source of all meaning, and the locus of all our suffering, our joy, our glory and shame. But this idea of immaterial souls, capable of defying the laws of physics, has outlived its credibility thanks to the advance of the natural sciences. (Dennett 2003, 5)

Contemporary neuroscience poses new challenges to the problem of free will. Consciousness is closely connected with the central nervous system; we know, for example, that certain brain damage affects consciousness. Furthermore, with advanced brain imaging such as functional MRI, studies have claimed to detect thoughts and decisions in the brain even before we become conscious of them. These studies further erode the concept of a sharp dichotomy between the body and the soul.

Muhammad Iqbal, an early 20th century Muslim philosopher critiques dualistic tendencies within Islamic thought and proposes a non-dualistic framework for understanding the problem of free will within Islam. Iqbal's analysis is, therefore, particularly relevant to this discussion.

According to Iqbal, dualism, as adopted by the dominant Ashari school of Muslim theology, has several problems.

1. The static view of substance does not serve any psychological interest. We do not think of elements of our conscious experience as qualities of a soul-substance. Iqbal writes, "Our conscious experience can give us no clue to the ego regarded as soul-substance, for by hypothesis the soul-substance does not reveal itself in experience" (Iqbal 2013, 81). Instead in Iqbal's view our conscious experience is exactly what makes up our self.

2. He points to Immanuel Kant's critique of the Cartesian dualism. Kant argues that the jump from Descartes "I think" to "I am a substance" is illegitimate and carries no proof.
3. The concept of the soul as indivisible does not prove indestructibility.
4. If the soul-substance is considered metaphysical then there is a whole host of other problems in trying to explain how it would interact with the physical body and soul.

Let us assume for the sake of argument that Dennett and Iqbal are correct. The question arises, to what degree is a theistic (specifically an Islamic) worldview dependent upon this type of dualism?

Islam & Free Will

I will argue in this paper, that in so far as soul/body dualism emerges from Islamic theological legacy, it does so due to practical considerations and not because of any intrinsic Islamic creedal commitment to such belief as essential dogma. The practical considerations were guided by the quest by the orthodox Ashari theologians to seek an intellectual framework that would reconcile essential Islamic beliefs with reason and philosophy.

Islamic intellectual history has a rich history of debate concerning the problem of free will. In fact, the very first formal theological dispute in Islam concerned the issue of free will and determinism (Blankinship 2008, 38).

Mu'tazilites

An early group of Muslims known as the Qadarites advocated for absolute free will, and this cause was then taken up by the more influential group of Muslim theologians known as the Mu'tazelites. Mu'tazilites stressed Divine justice. They held that good and evil are objective. In order for God to be just he had to punish those who do evil and reward those who do good. In order for this to be case, in order for Him to be just, humans must be able to act freely, so they can be held accountable for their actions. Therefore they argued that man is the creator of his own actions.

But this conflicted with the orthodox Ashari school, which thought that human free will in this sense would restrict the sovereign freedom of the creator. All acts are created and caused by God. God is all powerful and all knowing. Therefore man could not create his own actions. They promulgated a more nuanced theory of free will in which man acquired action but God remained the creator.

Aristotelians

On the other hand a group of early Muslim Aristotelian philosophers believed in a deterministic universe. They rationalized the existence of God, as the first cause, on the basis of reason. They argued if everything has a cause, and God is the first cause, then everything else must necessarily flow from that first cause resulting in a determined universe.

This led some of them, like Ibn Rushd, to reject free will. He writes, "Our actions occur according to a definite pattern… The determinate order of the internal and external causes is the decree and foreordination that God has prescribed for His creatures; that is the Preserved Tablet" (Ibn Rushd 1998, 189.)

This view also ran contrary to the more orthodox Ashari position which argued that a determined universe would deny God the ability to intervene within the world. A view championed by Ghazali.

Ghazali

The influence of Ghazali upon Islamic orthodoxy cannot be understated. Richard Frank, describes him as "the most important Sunni theologian at a crucial turning point in the history of orthodox Muslim theology" (1987–89, 274).

Ghazali demonstrated that necessary causation was a flawed doctrine, as Hume did many centuries later. He argued that cause and effect relationship cannot be proved; only a temporal succession of events is seen. In Hume's words there is no "causal glue" linking cause to effect.

He furthermore, more importantly, showed that necessary causation resulted in denying God freedom of will. If everything necessarily follows from previous events, then God would not be able to interfere with the workings of the Universe. For example, miracles would then be impossible. Therefore, for Al-Ghazali, the laws of causation were not necessarily true and could, in theory, be suspended by God at any given time. The primary purpose of Ghazali's argumentation during his time was to ensure God's freedom of will. His position on human free will was perhaps intentionally ambiguous, secondary to the theological debates of the time.

Professor Druart writes, "Whether or not Al-Ghazzali truly grants some agency to human beings is dubious, but he certainly wishes to grant it fully to God" (Druart 2005, 345).

In fact in both the case of the Mutazalites and the Aristotelians the primary concern of the orthodox Muslim theologians was preserving God's power and freedom of will.

The doctrine of human free will that emerges from this is simply a secondary consequence of that primary consideration.

Asha'rites

This emphasis upon granting God His attributes is a characteristic feature of Ashari theology which would become the dominant form of scholastic theology in Sunni Islam. Ghazali is generally identified with the Ashari school, although he may have had his differences. The Asha'rites adopted a dualistic picture of the human person one which consisted of a body and soul, with the soul serving as locus of personal identity. Iqbal summarizes their view when writes,

> To the Muslim school of theology of which Ghazali is the chief exponent [presumably Asharites], the ego is a simple, indivisible, and immutable soul-substance, entirely different from the group of our mental states and unaffected by the passage of time. (Iqbal 2013, 80)

This type of dualism served the purposes of the orthodox Asha'rite school in order to solve certain theological problems. However, as is being argued, Islam has no clear doctrinal commitment to it. Therefore, Islamic dualism differs in its nature when compared to the dualism which emerges out of Western Europe. It is worth pointing out the distinction in the substance of Islamic dualism and the context in which it arose.

Islamic &. Cartesian Dualism

The dualism from Islamic tradition differs from the Cartesian dualism which arises in the West. It is worth noting that Islamic dualism arises within the field of kalam, or rational/analytic/scholastic theology. This field is understood by Muslim theologians to be a tool or a dialectic method of dealing with intellectual challenges posed to Islamic doctrine. Kalam is generally not equated with Islamic doctrine or dogma itself. But is rather the art and science of defending this doctrine.

On its surface this appears to parallel the dualism we see develop in the western tradition, namely Cartesian dualism. But the type of dualism within Islamic scholastic theology, is not of the same category as the radical dualism found with Descartes. The dualism within Islamic theology does not arise of out radical skepticism, but rather out of pragmatic considerations. The primary pragmatic considerations were the rights and attributes of God. The Islamic concepts of free will and the soul are shaped based upon

concepts like the life hereafter, and preserving divine attributes like omnipotence and omniscience.

So instead of it being a theory of the human self, it is really a theory of God's interaction with humans. As Iqbal writes, "The unity of human consciousness which constitutes the center of human personality never really became a point of interest in the history of Muslim thought. The Mutakallimun regarded the soul as a mere accident which dies with the body and is resurrected and re-created on the day of judgment" (2013, 77).

According to this view, the soul is immaterial, indivisible, immortal and unchanging. This worked for the purposes of Islamic theological doctrines. The soul defined in this school had to be immaterial so it could be separated from the body. It had to be indivisible so it could not be destroyed, it had to be immortal so it could continue its life after the body died, and it had to be unchanging so that it could act as the anchor for personal identity. All of this was accomplished by creating the concept of the soul as a "substance."

This view is summarized by the celebrated contemporary Muslim Philosopher Dr. Naquib Al-Attas when he writes, "Man has a dual nature, he is both body and soul, he is at once physical being and spirit" (1995, 143). Ghazali himself repeatedly makes statements to the effect that "this subtle tenuous substance is the real essence of man" (Skellie 2010, 6).

A form of this dualistic soul picture becomes part of mainstream Islamic theology. But this happens not because there is an intrinsic Islamic basis for it but rather because it is a convenient tool in scholastic theology to deal with certain problems posed by other philosophers and theologians.

Origins of Islamic Dualism

To further examine how this type of dualism is different from Cartesian dualism we must examine its origin. This type of dualism originated with Ibn Sina and his floating man thought experiment. He asks his readers to imagine themselves suspended in air and isolated from all sensation without even sensory contact with their own bodies. He says the fact that we can imagine ourselves in this situation maintaining self-consciousness independently from the body implies that the idea of the self is not dependent on a physical thing. The soul therefore should be considered a primary given or a substance.

The type of dualism Ibn Sina introduces here is slightly more radical than that of Aristotle who regarded the soul as the form of the body, but Ibn Sina goes further and refers to it as a substance. Deborah Black writes,

> Ibn Sina does not reject the Aristotelian conception of the soul outright, but he upholds a form of soul-body dualism that is foreign to Aristotle… For Avicenna, the individual human soul is more than a physical entity and organizing principle for the body. It is a subsistent being in its own right, and a complete substance independent of any relation it has to the body. (Nasr 2002, 309-10)

Ibn Sina's experiment can be considered a precursor to Cartesian Dualism. However, there is an important difference. Ibn Sina's dualism is not as radical as that of Descartes. Descartes engaged in methodological doubt, doubting literally everything and concluding that only a metaphysical "I" could definitely exist. But Ibn Sina is not doubting the existence of the world or even of the body, these were already accepted as real in the Islamic worldview of his time.

Ibn Sina is only using the thought experiment as a way of securing for humanity an existence which is more than mere physical matter. He was using it to argue against a type of materialism that other philosophers were advocating. Despite his concept of dualism, however, Ibn Sina recognizes close ties between soul and body. He thinks of the body as an instrument of the Soul. He refers to it as a perfection of the body, or the captain of a ship or ruler of a city. In other words, he is still trying to maintain a link between the body and the soul. A link he ultimately cannot explain.

Ibn Sina was attempting to accommodate Islamic doctrines within a philosophical framework. Ghazali who was highly influenced by Ibn Sina, went a little further and critiqued philosophical methods when they contradicted Islamic doctrine while still attempting to maintain a rational worldview.

Although it seems Ghazali is subscribing to Ibn Sina's type of dualism by referring to the soul as a subtle tenuous substance, in fact there is more to the story that this. Even though he refers to the soul as a substance, he doesn't seem convinced that it is completely independent.

Jules Janssens, one of the foremost authorities on Ibn Sina in the world today writes,

> Whereas Ibn Sina justifies a sharp dualism between soul and body, this is far from the case in al-Ghazali. Indeed, he insists on the existence of a very special connection between the "subtle" heart and the "physical" heart. Referring to Sahl al-Tustari (d. 896) and his saying that the heart is the throne and the body the footstool, he points out that the relationship between them can be compared to that between God and His throne and footstool. However, al-Ghazali remains rather vague and

> admits that he consciously avoid offering any deeper explanation. In fact, he neither denies nor affirms a radical dualism between body and soul.... His designation of the "subtle intellect" as a particular expression for the seat of knowledge is of no real help in clarifying the issue. As to the subtle notion of 'spirit', al-Ghazali says nothing about it, except that it belongs to the "Lordly things," offering no clear explanation whatsoever. (2011, 619)

In Ghazali's famous text the Incoherence of the Philosophers, Ghazali actually argues against knowing that the soul-substance exists by pure reason, and is merely accepting it as part of religious law. The title of Discussion 18 is as follows: *On their inability to sustain a rational demonstration [proving] that the human soul is a self-subsistent spiritual substance that does not occupy space*.

In in he writes,

> We only want now to object to their claim of their knowing through rational demonstrations that the soul is a self-subsistent substance. .... We deny, however, their claim that reason alone indicates this, and that there is no need for the religious law. (Marmura 2002, 178)

For Ghazali the concept of the soul as a self-subsistent spiritual substance is consistent with revelation and serves the purposes of theology, so there is no problem with using it as a working theory. But even here he does not believe that this is inherently obvious via pure reason.

So his aim is primarily to preserve the importance of religious law, and he is okay with the contemporary concept of soul or self, which serves this purpose. We have seen that the concern of Muslim theologians was to secure the Islamic concept of God, and the concept of Man was simply a byproduct of ensuring that Gods attributes are preserved.

Unified Theory of Body & Soul

Putting this all in perspective, writing in the 20th Century, Iqbal then can make sense of an Islamic view on Free Will, which does not require such a sharp dualism between the body and the soul. Iqbal is arguing that given the problems with the classical formulation of the problem of free will and the soul a re-examination of the core Islamic doctrines reveals that indeed we are not committed to a dualistic picture of humans, but rather one that is not all that different from what Dennett has proposed.

According to this view the human person occupies a central role in the Islamic worldview. This is an underdeveloped concept in Islam's intellectual tradition, outside of the discipline of *tasawwuf* (Sufism). The Quran, he says, emphasizes the individuality and uniqueness of man. That man is chosen of God, that he is the representative of God on earth and he is a trustee of a free personality. To quote Iqbal again, he says, "It is surprising that the unity of human consciousness which constitutes the centre of human personality never really became a point of interest in the history of Muslim thought" (Iqbal 2013, 77).

Iqbal notes the difference in the word used by the Quran for creation of objects, *khalq*, with that used for the creation of the self, *Amr. Amr* is a directive command, it is a dynamic word. He writes, *Amr* "means, [that] the essential nature of the ego is directive, as it proceeds from the directive energy of God, though we do not know how Divine Amr functions as ego-unities." He writes, "the ego is present as a directive energy and is formed and disciplined by its own experience." And continues, "life of the ego is a tension caused by the ego invading the environment and the environment invading the ego" (Iqbal 2013, 82).

Iqbal writes: "Thus my real personality is not a thing; it is an act. My experience is only a series of acts, mutually referring to one another, and held together by the unity of directive purpose." According to this view the mind and body become one in action:

> When I take up a book from my table, my act is single and indivisible. It is impossible to draw a line of cleavage between the share of the body and that of the mind in this act. Somehow they must belong to the same system, and according to the Qur'an they do belong to the same system. "To Him belong Khalq (creation) and Amr (direction)" [7:54]. How is such a thing conceivable? We have seen that the body is not a thing situated in an absolute void; it is a system of events or acts. The system of experiences we call soul or ego is also a system of acts. This does not obliterate the distinction of the soul and body, it only brings them closer to each other. (Iqbal 2013, 84)

So according to Iqbal, we can discard thinking of selves as substances and focus on our real complete experience, which is best manifest in action. Professor Absar Ahmad explains this view:

> The self in its efficient aspect does not depend upon any obscure or hidden core, but depends upon what it does, has done, proposes to

> do, or is able to do. This self is revealed in its action; it reveals itself and constitutes itself by acting. It is nothing before acting, and nothing remains of it if experiences cease completely. One is not given a ready made self in this sense; one creates one's self daily by what one does, what one experiences. Our behavior is not an expression of our efficient self, but the very stuff which constitutes it. From the side of the efficient self, then, what holds experiences together, what gives us personality is not a substantial bond, but a functional one, a coordinated structure of activities. Being never a finished product, the efficient self is always in the making. It is formed throughout the course of its life. The efficient self, so to say, has no aboriginal nucleus of its own that exists prior to its action; it arises and takes on existence as it acts, as it undergoes experiences. (1986, 17)

In this view there is a distinct emphasis upon action. I think at this point Iqbal's view is not too different from that of Daniel Dennett on this point. Dennett writes,

> You have to distribute the moral agency around as well. You are not out of the loop; you are the loop. You are that large. You are not an extensionless point. What you do and what you are incorporates all of these things that happen and is not a completely separate thing from them.

Therefore, Islamic theology is not committed to dualism in the Cartesian sense. Moving away from a dualistic form of thinking could help solve at least some of the problems we typically associate with free will.

References

Druart, Thérèse-Anne. 2005. "Metaphysics." In *The Cambridge Companion to Arabic Philosophy*, edited by Pter Adamson and Richard C. Taylor, 327–48. Cambridge: Cambridge University Press.

Ahmad, Absar. 1986. *Concept of Self and Self Identity.* Lahore: Iqbal Academy.

Al-Ghazali, Abu Hamid. 2002. *The Incoherence of the Philosophers.* Translated by Michael Marmura. Salt Lake City: Brigham Young University.

Attas, Naquib. 1995. *Prolegomena to the Metaphysics of Islam*. Kuala Lumpur: International Institute of Islamic Thoughts and Civilization.

Blankinship, Khalid. 2008. "The Early Creed." In *The Cambridge Companion to Classical Islamic Theology,* edited by Tim Winter, 38-42. Cambridge: Cambridge University Press.

Dennett, Daniel C. 2003. *Freedom Evolves.* New York: Viking.

Frank, Richard. 1987–89. "Al-Ghazali's Use of Avicenna's Philosophy." *Revue des Etudes Islamiques* 55–57: 271–284.

Ibn Rushd. 2001. *Faith and Reason in Islam: Averoes' Exposition of Religious Arguments*. Translated by Ibrahim Najjar. Oxford: Oneworld.

Iqbal, Muhammad. 2013. *Reconstruction of Religious Thought in Islam*. Stanford: Stanford University Press.

Janssens, Jules. 2011. "Al-Ghazālī between Philosophy (*Falsafa*) and Sufism (*Tasawwuf*): His Complex Attitude in the *Marvels of the Heart* (*'Ajā'ib al-Qalb*) of the *Ihyā' 'Ulūm al-Dīn*." *Muslim World* 101 (4): 614–632.

Nasr, Sayyed Hossein. 2002. *Encyclopaedia of Islamic Philosophy*. Lahore: Suhail Academy.

Journal of Cognition and Neuroethics

Neuroscientific Free Will: Insights from the Thought of Juan Manuel Burgos and John Macmurray

James Beauregard
Rivier University

Biography

James Beauregard is a Lecturer in the psychology doctoral program at Rivier University, Nashua, New Hampshire where he teachers neuropsychology, biological bases of behavior and Aging. His research interests are in the fields of neuroethics and personalist philosophy including the intersection of these two areas as they impact our understandings of personhood. He is a member of the National Academy of Neuropsychology, British Personalist Forum, the American Catholic Philosophical Association, and the International Neuroethics Society.

Acknowledgements

I wish to thank Alan Ford and Simon Smith for their very helpful comments as this article was being prepared.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Beauregard, James. 2015. "Neuroscientific Free Will: Insights from the Thought of Juan Manuel Burgos and John Macmurray." *Journal of Cognition and Neuroethics* 3 (1): 13–37.

Neuroscientific Free Will: Insights from the Thought of Juan Manuel Burgos and John Macmurray

James Beauregard

Abstract
Neuroscience has moved to the forefront of the free will debate, introducing new ideas, new questions and new problems regarding Free Will. This paper examines our neuroscientific and philosophical notions of free will and suggests that the principal problem in the debate is that the free will question has been inadequately formulated. The neuroscience debate has come to be grounded in an empirical neuroscientific position with its materialist/deterministic presuppositions, and in the outmoded position of substance dualism that has bequeathed us the mind/body problem. A reformulation of the question is presented drawing on the thought of Juan Manuel Burgos and John Macmurray in the tradition of philosophical personalism. Free will, when examined from the field of the personal, and in categories specific to human persons, becomes a human activity, rather than a merely physical capacity or faculty. From this perspective, it becomes possible to address free will in the context of normal development, and to ask questions regarding limitations on free will, including contemporary understandings of the impact of neurologic and psychiatric illness.

Keywords
Free will, personalism, Juan Manuel Burgos, John Macmurray, neuroscience, neuroethics

I. Neuroscience and Free Will

Neuroscience has made its presence felt in the Free Will debate in two ways. First, and most pervasively, through the methodology and general worldview of science present in the wider culture, as science investigates brain function and shapes our views of self. Secondly, and most dramatically, neuroscience entered the Free Will debate through the work of neuroscientist Benjamin Libet in his empirical studies of conscious intention to act, studies that have generated widespread discussion and conflicting interpretation.

The purpose of this essay is to address the question of free will at the structural level and to question of whether our current conceptual approaches are adequate to the task. I will suggest that the "problem" of free will is one that is largely of our own making and is the result of a process of abstraction that limits our ability to address the question. I will also suggest that a broadening of this currently over-limited conceptual architecture

would be helpful in seeing the larger problem, and moving toward a broader and deeper discussion of free will that takes into account the full complexity of the human person.

Neuroscience

Neuroscience has deepened our understanding of brain function as a complement to discussions of Free Will. It has also been problematic in terms of the underlying assumptions of its worldview. The methods of neuroscience are those of the wider scientific enterprise, grounded in the experimental method and the examination of observable phenomena.[1] Science in the modern era limits itself to the observable, physical world, and thus on the physical/material and organic aspects of the person, a focus that has brought about great advances in science and medicine, but at the same time continues to operate within a dualistic and at times monolithic world view that creates difficulties when attempting to think about Free Will, forcing us to ask the question, "How our mind and body connected?"

The Necessary Consequence of Dualism

This dualism arose as philosophy shifted from a theocentric to an anthropocentric enterprise. Cartesian notions of mental substance and physical substance as discrete and fundamentally different aspects of reality have left us with the struggle of trying to describe how the two interact. Neuroscience has, in practice, embraced and reduced this dualism in its discussions of mind and body, and attempted to describe these two conceptually distinct entities in terms of neurobiology.[2] I would like to suggest that the problem of Free Will, within this context, is a problem largely of our own making.

---

1. See, for example, Gazzaniga et al (2009) *Cognitive Neuroscience: The Biology of the Mind,*. 3rd ed. New York: W.W. Norton & Co., especially Chapter 4, "Methods of Cognitive Neuroscience."

2. Neurologist Hal Blumenfeld captures this debate: "Where is the mind and what is the mind? These questions have haunted scientists and philosophers throughout human history. Although we cannot yet answer these questions with certainty, investigation for he nervous system allows at least tentative conjectures in this realm. Although some would argue otherwise, the burden of evidence currently available suggests that the mind is manifested through ordinary physical processes located within the body. Note that these first two fundamental conjectures about where the mind is (in the body) and what the mind is (normal physical processes) remain hypotheses, perhaps with growing evidence in their favor, yet remaining unproven nonetheless. See his *Neuroanatomy Through Clinical Cases,* 2nd ed., "A Simple Working Model of the Mind," 973 ff.

Free Will in the Field of the Personal

For Free Will there is a way through the problem of dualism that can be found in the philosophy of Personalism. In examining this path, I will look to the work of the Personalist tradition in general, and specifically to the work of two Personalist philosophers, John Macmurray, writing in Scotland and England through much of the 20th century, and Juan Manuel Burgos writing in Spain today. Both philosophers offer a diagnosis of the problem of how we think about Free Will, and both offer some solutions.

If we follow the work of these two philosophers, it becomes possible to see the conceptual difficulties in many of the current formulations of the problem of Free Will to reformulate them in a more comprehensive way. These philosophers address some problematic philosophical ideas that have endured in the Free Will discussion, and seek to reformulate a number of questions in a more comprehensive manner.

II. Personalism

The Personalist philosophical tradition has deep roots in the history of philosophy, but took its contemporary form during and after the two World Wars. Personalism steps outside of the Cartesian thought world and offers an opportunity to consider the question of Free Will in a manner that is grounded in persons adequately envisioned. For these philosophers, the question of Free Will is first and fundamentally a question of Persons. Personalism originates in a shift in perspective, moving from questions of being to questions of doing, from questions of substance to questions of action-in-relation.

Personalism is not a single philosophical school, but rather a "worldview" defined by some central ideas held in common by many philosophers working in the tradition. It is grounded in "the general affirmation of the centrality of the person for philosophical thought" (Williams and Bengtsson 2014, 2). Personalism asserts "the person is the key in the search for self – knowledge, for correct insight into reality, and for the place of persons in it" (Buford 2011, 1). In practice, this means that "Personalists believe that the human person should be the ontological and epistemological starting point of philosophical reflection. Their concern is to investigate the experience, the status and the dignity of the human being as person, and regard this as the starting – point for all subsequent philosophical analysis" (Williams and Bengtsson 2014, 3). British philosopher Richard Allen has given us a succinct description of the Personalist tradition:

> 'Personalism' is a distinctive way of thinking, and not only in philosophy but also in theology, history, sociology and psychology, which stresses the distinctiveness, unique value, freedom and responsibility of

> personal existence, and seeks to articulate and apply the categories and conceptions uniquely appropriate to persons, and not just those applicable to animals, organisms and merely physical entities, nor the barren and abstract ones of formal logic. (Allen 2013, 2)[3]

Personalism encompasses a broad range of philosophical ideas and methods; historically, it has never associated itself with one specific or dominant methodology. At the same time, with Person as its central *topos* (Buford 2009), Williams and Bengtsson (2014) have identified five common characteristics of Personalist thought:

1. The Person/Non-person distinction: Personalism holds that there is radical difference between persons and nonpersons and that there is a character of irreducibility to the human person. For Personalists, the difference between human beings on the one hand, and animals and inanimate objects on the other is one of kind, not of degree.
2. The dignity of the human person: The ontological distinction between persons and nonpersons yields ethical consequences that include affirming the dignity and inherent value of persons, recognizing a difference between "someone" and "something" which touches on issues of justice, political decision-making, and life in community.
3. Interiority and subjectivity: Personalism acknowledges the interior/subjective nature of person in the unity of self-awareness and consciousness, in the human experience of self as both subject and object of activity, not reducible to the material world. In the Personalist tradition, this subjectivity encompasses "interiority, freedom, and personal autonomy" (Williams and Bengtsson 2014, 31). "The human being must be treated as a subject, must be understood in terms of the modern view of specifically human subjectivity as determined by consciousness" (Williams and Bengtsson 2014, 33).
4. Self-determination: For Personalism, this is the domain of Free Will in action, where persons interact with the world not in a causally predetermined manner, but rather act out of an inner subjectivity capable of action and self-determination. For the Personalist, human action is creative, causal, and determines both the external world and the actor. Action and self-determination occurs in the domain of the moral life where good or bad actions are sought and

3. Allen's brief, focused essay is in the introduction to *Three British Personalist Philosophers*, which gives overviews of the work of Michael Polanyi, John Macmurray and Austin Farrer. It is available through the British Personalist forum, www.britishpersonalistforum.org.uk.

performed and where persons form themselves into morally good or bad human beings.

5. Relationality: The various Personalist traditions emphasize the human person's nature as a social/relational being. Persons never exist in isolation, and becoming a person is an integrally social activity that happens across the lifespan. For the Personalist, human beings flourish most fully only in relation with others. This aspect of personhood was strongly influenced by mid-20th century Personalist philosophy seeking a middle way through the extreme visions of Totalitarianisms of the right and the left that saw the individual as subsumed by and subordinate to the state on the one hand, and an extreme individualism which viewed the human person without need for relation to others. For the Personalist, humans are naturally social, naturally form societies and maintain them intentionally in the service of human flourishing.

III. Diagnosing the Problem

Juan Manuel Burgos and John Macmurray, both operate within this broadly Personalist tradition. Both writers illustrate that much of the "problem" of free will is one of our own making, deeply rooted in our habits of thought formulated at the beginning of the scientific revolution in the 17th century and in the concurrent rise of the modern philosophical period. Both authors see much of the difficulty in contemporary thought as grounded in a misconception of persons, and both propose a transformation in our thinking to a more holistic vision of person that adequately addresses the problem of Free Will.

Juan Manuel Burgos and the Classical Philosophical Heritage

Juan Manuel Burgos, currently Professor at the University San Pablo CEU (Madrid) has written extensively in the Personalist tradition.[4] He asserts that in order to adequately describe persons, an integration of the best of classical and contemporary philosophical concepts are necessary. At the same time, this process necessitates eschewing philosophies of person that attempt to describe what a person is indirectly.

He refers to this problem of indirect description as our "Greek ballast." In classical Greek philosophical tradition, the *Cosmos* encompassed all things, the natural world,

4. His works include *Introdución al personalismo* (2012, currently being translated into English), *Antropología: una guía para la existencia* (2009), *Reconstruir la persona. Ensayos personalistas* (2009) and *Repenser la naturaleza humana* (2007).

human beings, the gods. Within this conception of the universe, stood the great biologist of the ancient world, Aristotle. In his classification scheme, he placed persons under the broad category of animals, and sought to distinguish human beings by describing them as rational animals. While capturing an aspect of what it means to be a person, the term created conceptual difficulties:

> By the Greek ballast I mean the tendency, born in Greek philosophy, describing man by applying slight modifications to philosophical notions designed for objects or animals, with the result that what is specifically human, what constitutes man as a person is obscured or may even disappear. (Burgos 2013)[5]

The Problem of Categories

A large part of our difficulty in understanding the nature of personhood, for Burgos, is the conceptual error of trying to understand human beings primarily through animal categories that are modified in some way to reflect specific aspects of human beings (e.g., "rational animal").

To be able to understand and talk about persons, he argues, it is necessary to put aside this conceptual baggage and to develop an anthropology that begins with considering what is specifically and uniquely human. If we begin with persons, rather than animals, it becomes possible to construct an anthropology that reflects what is uniquely human, rather than trying to shoehorn persons into categories the do not fully capture who we are.

Burgos describes this process in the history of philosophy from the early 20th century forward as "the personalist turn" (Burgos 2013, 8) reflected in all of the philosophers outlined above, the process by which human beings are approached not from extrinsic categories, but from categories exclusive to persons (Burgos 2013, 5). "The indiscriminate application of general concepts, such as the four causes, accidents, substance, to any kind of being, is a simplification that is not justified given the complexity of reality, and leads to confusion and a poor understanding of the issues involved" (Burgos 2013, 7). The most fundamental conceptual shift is moving from asking "what" is a person to asking "who" a person is.

5. Burgos expands on these ideas in *Repensar la naturaleza humana* (2007), 59–63, "The Greek ballast and the problem of enlargement" (El lastre griego y el problema de la amplicación).

Another consequence of this conceptual ballast is the mind-body problem, with roots that can be traced back into ancient Greek philosophical tradition of soul and body, adopted in the Middle Ages, where e.g. Thomas Aquinas described the soul as the substantial form of the body, linked in some way to the body as its animating principle. The problem was immeasurably deepened by Descartes' radical doubt that resulted in focusing on thought and the theoretical to the detriment of the physical and the practical.

In attempting to reformulate our understanding of persons anew, Burgos attempts to articulate a notion of person in categories specific to persons. He posits a tri-dimensional structure of person, described in three integrated "levels" of body, psyche and spirit, a structure which stands against dualism allows for a more sophisticated philosophical anthropology (Burgos 2013, 9). An important aspect of this process involves a rehabilitation of emotion, seeing it not as something interfering with rationality, not an expression of the irrational in us, but as a unique and integral dimension of persons:

> To overcome this vision, something necessary to achieve an integral and apology, we have to hold two points. The first one is the originality of emotions, i.e., it's radical difference from human knowledge and from human dynamisms. And, to do so, we might understand them as the way the person is present to himself in his subjectivity (vivincia de sí). To feel is different from to know and from will. To feel is to feel, the same way to see is nothing more or less than to see. It's a *primary* anthropological dimension. In second place, we have to be conscious that this anthropological trait is present in all the tri-dimensional anthropological structure of the person: body, psyche and spirit. There are bodily feelings, that is to say, the way we feel the body; there are emotions in the psychic level; and there are also spiritual feelings, which give reason of some of our deepest personal experiences like the relations with our beloved ones. (2013, 10)

Feelings, then are intimately connected with our subjectivity and our consciousness, our experience of ourselves as persons: "to live himself (*vivirse al sí mismo*); to possess a unique personal world is an essential trait of human being, the trait in fact, which transforms him into a 'who'" (Burgos 2013, 10).

Persons in Relation

A capacity for interpersonal relationships is also central to Burgos' philosophical anthropology, in keeping with the wider Personalist tradition of viewing persons in relation rather than as isolated selves. We are born into a web of relationships necessary for our development as persons; absent this relational world persons cannot develop, or even survive. We are social by nature and live in a world that begins with relationships with parents, in the context of immediate and extended family, and broadens into a larger world of "I and Thou."[6]

At the same time, persons do not exist in a world of relational abstraction. For Burgos, body is a dimension of person, removed from dualistic notions of mind-body that create the need to try to explain how the two interact. Our bodies are not something we *have* or something we *inhabit*, but rather a *dimension of who we are*, the somatic dimension of personhood. In this vision, for example, human sexuality is a personal, normal and integral dimension of personhood. While our sexuality is rooted in our biology, biological processes alone cannot fully apprehend or express the richness of sexuality touching on all dimensions of the person (Burgos 2013, 12). Continuing this relational notion of person, Burgos describes persons as naturally social, and human society not as a necessary evil, but a natural expression of personal activity. Persons are seen as the center of society, the standard by which society ought to be organized.

Lastly, these aspects of personhood raise the questions of how persons relate to one another. In the Personalist vision, this means moving beyond a static philosophy of being to a dynamic philosophy of action and interaction that is integrated and respectful of persons in all their dimensions. This means moving beyond ancient and medieval philosophical traditions focusing on intellect that led to the modern philosophical focus on epistemology, logic and language to a philosophy grounded in action:

> Now, Praxis, understood as the medium in which man expresses and transforms himself became central; and people also realized that man will be understood really only fully if all the dimensions of its [man's] activity are also fully understood. This new orientation allowed personalism to deal with many areas that scholastic tradition had neglected like work, aesthetics, economy, social and political philosophy, and so on." (Burgos 2013, 11)

6. Burgos is drawing here on the tradition of dialogical personalism, most famously and enduringly expressed by Martin Buber in his *I and Thou* (*Ich und Du*).

For Burgos, as for many other personalist philosophers, the only proper relation between persons individually and in society, is love. "Personalism emphasizes the priority of love as a guiding element of human activity in so far as it gives meaning to life and to interpersonal relationships. A life without love, in which someone had not been loved or could not have love, would certainly be a life radically inhuman and incomplete" (Burgos 2013, 11).

J.M. Burgos on Free Will

It is in the context of philosophy of action-in-relation that Burgos specifically addresses the question of Free Will, as an integral dimension and activity of persons. Within a philosophy of action, we do not *have* Free Will; rather the *do* Free Will as an integral aspect of who we are. Freedom, for Burgos is a "deep feature" of human dynamism that can be obscured if one approaches persons in categories other than those specific to person, including the physical and the biological, grounded as they are in theories of cause and effect. To approach persons in categories specific to persons allows us to observe human dynamism and free will and action. It is only when we prescind from categories specific to persons and focus our attention on physical and biological analogies that free will become a problem. Operating, rather, in a conceptual world that focuses on categories specific to person, freedom is specific to persons, and Free Will is seen "specifically as self-determination."

Personalism, then, has a fundamentally ethical nature touching on our relationships with ourselves and with others directly and at increasingly complex levels of family, community and society. In order to articulate a moral vision in this context, it is necessary to begin with persons fully conceived in categories specific to persons, a necessary process without which persons cannot be fully apprehended and understood.

John Macmurray and the Field of the Personal

Scottish born John Macmurray (1891–1976) wrote and taught for many years in the Personalist tradition; his work touches on such topics as the social nature of the person and the need to move from a philosophy of being to a philosophy of action.[7]

7. His works include *The Self as Agent* (1991), *Persons in Relation* (1991) and *Reason and Emotion* (1992).

The Concept of "Field"

Macmurray's basic philosophical enterprise emerged in his analysis of the modern philosophical period, as he attempted to articulate the nature of the human person individually, in action, and in relation, distinct from the physical and biological analogies of person that developed from the 17th through the 19th centuries. To do this, Macmurray uses the concept of "Form" or "Field," by which he meant a conceptual architecture used to understand the human person.

For Macmurray, every age has its own central philosophical questions, and as transformations in society occur, old questions and problems diminish in importance or fall away and new questions arise. He located modern philosophy's origins in the period of the early scientific revolution, and in the turn to a focus on the individual. For Macmurray, Descartes effected a disruption in a unified vision of person that existed prior to the modern philosophical period:

> Modern philosophy is characteristically egocentric. I mean no more than this: that firstly, it takes the Self as its starting point, and not God, or the world of the community; and that, secondly, there is an individual in isolation, and ego or 'I', never a 'thou'. This is shown by the fact that there can arise the question, 'How does this Self know that other selves exist?' Further, the Self so premised is a thinker in search of knowledge. It is conceived as the Subject; the correlate in experience of the object presented for cognition. Philosophy, then, as distinct from Science, is concerned with the formal characters of the processes, activities or constructions in and through which the object is theoretically determined. And since the Self is an element, in some sense, of the world presented for knowing, it must be determined through the same form as every other object." (Macmurray 1957, 31)

Stage 1: The Field of the Material

Macmurray characterized the philosophical era that runs from 1600 to the present as falling into three broad stages. The first of these stages ran from the beginnings of the scientific revolution through the mid-1800s. In the sciences it was typified in the discipline of physics and the work of Isaac Newton. The development of the science of physics was grounded in a physical/material vision of the world, described in mathematical methodology, deterministic in nature, that saw the fundamental mode of activity in the universe as that of cause and effect. The overarching philosopher of the age

was René Descartes, in a philosophical process running from Descartes to David Hume, leaving us with a seemingly inescapable dualism of mind and matter, a world in which there were two types of "substances," one mental and one physical. This initial dualism devolved to numerous positions in the philosophical tradition ranging from absolute idealism to absolute materialism, the universe reduced either to matter or to mind. This left philosophy with the unenviable task of trying to derive a vision of person from one or the other of these polarities. Science set up camp in the materialist vision with its paradigm of cause and effect, which Murray termed this "Field of the Mechanical" (Macmurray 1957, 13), a view of the world is composed of matter, describable through mathematics, deterministic in nature, that sees human beings as purely material and explainable in these terms. It is a field that is fundamentally impersonal, in which it is impossible to conceive of Free Will.

Stage 2: The Field of the Organic

The next stage Murray characterizes involves a reaction to the world view of physics, an era that saw the development of the science of biology, and in a series of reactions, both philosophical and artistic, to purely mechanistic notions of the workings of the universe. If the central figure in the Field of the physical/material/mechanistic was Isaac Newton, the central figure of the field of the organic was Charles Darwin. Out of his theory of evolution developed the field of evolutionary biology, and the attempt to understand persons through biological/organic categories and analogies. This involved a return to ancient Greek notions of person, viewed in animal categories. This developed, in the late 19th century, to a view of society as a developing and evolving organism, one of whose consequences was the development of Social Darwinism, in which not only individuals but also ethnic groups were engaged in a struggle for survival. In its most negative aspects it was seen in the eugenics movement of the late 1800s, and taken over into the political realm in National Socialist racial policy of the 1930's and 40's.

It is here, in the Field of the Organic, that consciousness arises in the animal world. The Field of the Organic continues to operate in an essentially deterministic mode, now conceptualized as the stimulus – response of biological organisms engaged in the process of adaptation to environment. It is still a world that Murray would characterize as "impersonal":

> Greek tradition has been strongly reinforced by the organic philosophies of the 19th century in the development of evolutionary biology. This in turn led to the attempt to create evolutionary sciences

> in the human field, particularly in its social aspect. The general result of these converging cultural activities – the romantic movement, the organic philosophies idealist, realist and evolutionary science, – was that contemporary thought about human behavior, individual and social, became saturated with biological metaphors, and molded itself to the requirements of an organic analogy. It became the common idiom to talk of ourselves as organisms and of our societies as organic structure; to refer to the history of society as an evolutionary process and to account for all human actions as an adaptation to environment. (Macmurray 1991a, 45)

In the end, the attempt to understand human nature through the organic analogy is, for Macmurray, a structural error: "a categorical misconception is a misconception of one's own nature… If, however, the error lies in our conception of our own nature, it must affect all our action, for we shall misconceive our own reality by appearing to ourselves to be what we are not, or not to be what we are (Macmurray 1961, 149).

Stage 3: The Field of the Personal

As noted earlier, Macmurray was convinced that previous philosophical traditions had run their course and ended in bankruptcy. Attempts to understand persons foundered in models and analyses that were materialist or biological in nature. Macmurray's response to the situation was the development of his own Personalist philosophy. The Field of the Material and the Field of the Organic had failed us, in his view; what was needed was a turn to the Field of the Personal.

In making this turn, Macmurray realized that he was entering into uncharted territory. In 1929, he wrote to a friend,

> it seems to me that we have not yet begun the effort to understand the Personal at all, and that we don't yet have the logical apparatus to do it. We know persons and personal activities – nothing better: but when we try to understand them or express them we do so always buy him personal analogies – drawn from the physical or the organic world. Now the logical structure of the personal is radically different from either of these. (cited in Macmurray 1992, xi)

Macmurray wrote that for a solution to the problem it was necessary to step outside of the Cartesian system altogether, to move from thought to action:

> And since the effect of transferring our point of view from the 'I think' to the 'I do' is to overcome the dualism which is inseparable from the theoretical standpoint, the dualism of a rational and empirical self disappears. There is no longer any need to isolate the two aspects of unity and difference in an antinomy of shared identity and sheer difference. A personal being is at once subject and object; but he is both because he is primarily agent. As subject he is 'I', as object he is 'You', since the 'You' as always 'the Other'. The unity the personal is, then, to be sought in the community of the 'You and I', and since persons are agents, this community is not merely a matter-of-fact, but also matter of intention. (Macmurray 1991a, 27)

For Macmurray, to move from the Fields of the Material and Organic to the Field of the Personal is to move into a conceptual architecture that takes Person as Agent, as relational, as the starting point of philosophical thinking, to develop a philosophy of action and intention, and to include the Material in the Organic conceptions within the broader concept of Person for a comprehensive reintegration of the unity of persons. To conceive of persons, particularly persons as agents, is to move beyond the limitations of determinism and into the realm of freedom. For Macmurray, freedom is "the capacity to act" (Macmurray 1991a, 98). Persons, then, are subjects, objects and agents simultaneously. The Personalist vision is inclusive rather than exclusive, encompassing the specifically human in a world in which consciousness and action are integrated in "a unity of movement and knowledge."

To enter into the Field of the Personal is to adopt a way of seeing and a mode of activity that is inclusive; this is a crucial notion in that it allows us the place the previous movements of the modern period in context, to see the necessity of their inclusion in the concept of person, but also to realize that we are not limited to those concepts: "That the concept of 'a person' is inclusive of the concept of 'an organism, as the concept of 'an organism' is inclusive of the concept of 'material body'" (Macmurray 1957, 118).[8]

8. It should be noted that there is debate within the personalist philosophical community about the extent to which interpersonal *relations* create and define persons, either completely or partially. While a purely relational/functional understanding of persons runs the risk of assuming "no relation = no person" a purely individualist notion of person centered in the body and its operations runs the risk of materialism and dualism; a balance is needed between the person and the person in relation. Also, in a critique of Macmurray's philosophy, Robin Downie wrote, "Like Macmurray I hold that the ideal of community' is important, but unlike Macmurray I wish to avoid the tyranny of the personal. Macmurray says that our identity is constituted by our personal relationships. I wish to hold the more modest thesis that our

Attaining the Field of the Personal, it becomes possible to examine the strengths and limitations of the two earlier stages of the modern philosophical period, and to come to see them as necessary but not sufficient for understanding human beings. Macmurray argues that our very understanding of these previous stages, the material world and the organic world, stems from the vision that takes place within the Personal world by a process of abstraction and limitation of attention. He argues it only *as persons* can we conceive of concepts of materialism and organicity. In terms of our conceptualizations of a material and biological world,

> It was assumed, and still is assumed in many quarters, that this way of conceiving human life is scientific and empirical and therefore the truth about us. It is in fact not empirical; it is a priori and analogical. Consequently it is not, in the strict sense, even scientific. For this concept, in the categories of understanding which go with it, were not discovered by a patient unbiased examination of the facts of human activity. They were discovered, at best, through an empirical and scientific study of the effects of plant and animal life. They were applied by analogy to the human field on the a priori assumption that human life must exhibit the same structure. (Macmurray 1991a, 45–46)

The Field of the Personal, provides us the key to unlocking the Problem of Free Will and the way out of its insolubility. Throughout much of the Western philosophical tradition in general, and in the philosophy of the modern period in particular, we have attempted to think about Free Will from the Fields of the Material and the Organic, and thus have made it an insoluble problem. It is a problem of our own making because we have been looking in the wrong place: *Free Will resides not in the Material or Organic domains, but in the Field of the Personal*.

When we begin to look to the Field of the Personal our problem of Free Will dissolves, the material and organic aspects of human nature fall into place, and we can recognize free will as an integrated human activity, an activity of the whole Person.

---

identity is partly constituted by our personal relationships, but is also influenced by such factors as the environment, the arts, animals and so on. Human beings are complicated and the relationships which constitute our identity and make us flourish are correspondingly diverse." See Downie, Robin. "Personal and Impersonal Relationships," in, D. Ferguson and N. Dower, eds., *John Macmurray: Critical Perspectives*. New York: Peter Lang Publishing, 2002, 131.

IV. Free Will in the Field of the Personal

Macmurray's philosophy gives us a way of understanding the nature of Free Will and to come to see it not as a problem but as a dimension of persons. When we think about the world, we think about it *as Persons*. In fact and in practice we operate in a unified way. The personal includes the Material and the Organic, but can only be understood, as, Burgos also noted, by examining categories specific and unique to persons, categories that do not exist in the material and biological worlds. These categories involve our existence as persons in a personal world of interrelation, in which emotion, reason and knowledge exist in the context of human subjectivity expressed through our bodies, not bodies that we *have*, but bodies that we *are*. These dimensions of our personhood exist in interpersonal relation, and are formed and developed through relationships with other persons from the very beginnings of life; to think of Persons outside of relationships with others is a virtual contradiction in terms; we are born into relations, develop within them, and live out our lives in relations with others. Our bodies are "the somatic dimension of the person "separable only theoretically and in abstraction. There are no persons without bodies and "There is no real body without a person" (Burgos 2013, 12). Our sexuality, for example, is not something that we do, not a merely biological activity, but touches "the very constitution of the subject. The person, in fact, not only possesses a male or female biology, but is a man or woman, a male or female person, because sexuality touches all human structures giving them a peculiar character" that does not exist in the animal kingdom (Burgos 2013, 12). Persons are both subjects and agents, who come to know each other not in thought, but in integrated action, as doers, as agents in personal relation.

The Problem of Abstraction

In this sense, the problem of free will reveals, itself as a problem of abstraction. Persons are integrated wholes, separable in theory, but not in practice. "In practice we understand any form of behavior better the closer it is to our room. All human knowledge is necessarily anthropomorphic, for the simple reason that we are human beings" (Macmurray 1957, 116). As Macmurray describes it, we understand the material and biological words by beginning with our personal knowledge of ourselves and abstracting from them. The Organic and biological worlds exist when we abstract, when we remove, the Personal. The Material world exists when we begin with the Personal and abstract both the Personal and the Organic, leaving the world of matter in motion, following deterministic laws. In removing the Personal, we remove the domain of the

specifically human; rationality, freedom, self-determination and self-transcendence. What is left is a deterministic world, either living or inorganic, toward which we direct our attention in a limited way, attending not to the personal, but to what remains when the Field of the Personal, or the Field of the Personal and Organic are removed. It is important to remember that these states exist in theory, in abstraction, but not in reality. The danger arises when the parts are reified or inflated to become a whole, when the material or the organic are enlarged and equated with the whole of reality. When this happens, all we have left are cause-and-effect, stimulus and response, causal determinism that makes it impossible even to conceive of Free Will. This process of abstraction is one in which materialist philosophies and conceptions of evolutionary biology understood exclusively in notions of genetics operate. They share a common, and a fundamental error, the mistaking of the part for the whole. In these domains there is no possibility of adequately conceiving the concept of Free Will and one is left with the necessity of stating that there is no such thing. The great irony in this process is that it is ultimately self-defeating. In a world of matter only, all that can exist is matter in motion, with all activity predetermined by previous activity in an infinite regress, the physical world in which meaning and freedom cannot exist. However, order to assert a materialist philosophy in any form, it is necessary for the materialist to step outside of materialism and speak from the Field of the Personal, the world in which he or she was formed by others, taught to speak, nurtured and developed, in order to deny that the Field of the Personal exists. To be a materialist who denies the possibility of free will is to do so as a Person exercising their Free Will in the act of denial.

Free Will exists, it operates in us robustly, it is easily recognized from person to person, recognized in others and within ourselves; What we need to do is to look in the right place, the Field of the Personal, and avoid the well-worn habits of mind that would abstract us from this, leaving us trapped in a materialist or dualist world.

V. Neuroethics and the Field of the Personal

What would Neuroethics look like if conceived from the Field of the Personal? We would recognize free will as an integral aspect of the human person exercised across the lifespan, in a more limited, developing way in childhood, reaching full expression in adulthood and continuing robustly (barring illness or injury) into old age. Free will in this context would necessarily be seen as an ethical, as a moral activity since it involves interactions between persons, human flourishing, and the right use of the natural world.

The construction of a Neuroethics in this vision would begin not with brains, but with Persons adequately understood, persons for whom a somatic dimension is a part of that personhood, a somatic dimension that includes body and brain, and from which some common moral norms can be derived. A neuroethical vision so derived would attend to human persons in all aspects, material (a descriptive project), the organic (a descriptive project), and the Personal (a descriptive and prescriptive project).

Neuroethics, Free Will and Persons Adequately Considered

A neuroethics of Free Will thus conceived would begin by examining the normal development of our rational and affective capacities, recognizing the originality of both, and their deep interconnection in the process of decision-making. It would examine the exercise of free will across the lifespan as exercised by normal, healthy individuals and apply that knowledge to the various domains with which Neuroethics is concerned, some of these being personal autonomy, consent to participation in research and medical care, and personal responsibility for our actions freely conceived and carried out, in this context including considerations of personal responsibility and our justice system. While the Material and Organic visions of person could easily do away with the need for legal system, such a system remains essential when viewed from the perspective of freely acting, responsible Persons.

Limitations of Free Will

With these conceptions in mind, and only after this has been attained, can one adequately consider limitations of Free Will. Reason and freedom were the classical complements of human action and moral responsibility. Classically there are several things thought to be impediments to freedom, to free action and thus limitations on moral responsibility including ignorance, fear, coercion and passion.

The contribution of neuroscience to this picture is the deepening understanding of the structure and functioning of the human brain, including normal development and function, and conditions that can place limitations on that function including neurologic injury (e.g., traumatic brain injury) in neurologic and psychiatric illness (e.g., dementing illnesses, schizophrenia, bipolar disorder, anxiety and depression, hallucinations, delusions, etc.).

A Neuroethics moving from the Field of the Personal would not limit itself to the individual, or to the brain function of individual, but would consider the whole person in the context of our relational world. Issues of both individual and common good would

be considered, impacting on the choices made with new technologies, the allocation of resources in distributive justice, in the manner in which our scientific knowledge is put to use in a variety of circumstances.

The guiding value in this vision would be the good of the human person fully and adequately conceived, including categories specific to person such as reason, freedom, transcendence, and the human capacities for rationality, relationality and sexuality, the person as subject and as agent.

I would like to conclude with a brief mention of several of the major domains of Neuroethical research as they might be impacted by the philosophical positions of Macmurray and Burgos presented here.

Cognitive Enhancement

A Neuroethics grounded in personhood which encompasses the physical and the organic but does not limit itself to these might examine issues of cognitive enhancement empirically at the individual level, providing a realistic appraisal of its effectives (or lack thereof), but also see it in the broader context of societal issues including distributive justice and our understandings of health and illness; it would make recommendations based on a comprehensive review of the data, not just neurologically, but as it would impact our education system, healthcare, the workplace. It would not begin with cognitive enhancement as a given or as inevitable, but would examine these questions from the notions of both personal and common good.

Free Will, Responsibility and the Justice System

A Personalist Neuroethics would recognize the existence and activity of Free Will, and with that the reality and necessity of personal responsibility. It would not limit itself to the organic or the material, and thus would not become trapped in concepts of determinism that raise basic questions about whether personal responsibility can even exist, or whether there is a need for a justice system. It would also recognize legitimate limitations on freedom and responsibility in terms of neurologic injury or dysfunction impacting on a higher-level cognition to processes. It would work to educate judges, attorneys, and juries about the accurate and proper use of neuroimaging data, seeing it in the context of personhood and not falling into the trap of equating discrete neurologic lesions with personal responsibility or the lack thereof. It would promote the sustained and serious consideration of ethical uses of neuroimaging technologies in the courtroom, in the medical field, and in national security.

Neuroethics and Capacity/Competency

A Personalist Neuroethics would address issues of capacity and competency by beginning with an adequate, comprehensive vision of Person and seek to articulate those conditions under which personal autonomy is diminished or should be limited for an individual's safety. It would do so through consideration of the categories unique to persons, the nature of limitations on those categories through injury or illness, in the threshold below which a person would not be considered able to make decisions for themselves.

Neuroethics and Medicine

This has been, and likely will continue to be the most contentious arena in which Neuroethics has input, an area which generates the strongest of feelings across political and religious spectrums, touching on issues of when human life begins, stem cell research, abortion, assisted suicide and euthanasia, and in decisions about medical care in cases of persistent vegetative state and brain death, and treatment decisions at the end of life.

The Beginning of Life

Within the domain of neuroscience, various ideas have been put forth as determinants of when personhood begins. Each of these conclusions is built on an assumption of dis-integration. By this I mean that inherent in each of these attempts a false assumption that the different dimensions a person can be separated out not only abstractly for understanding, but in reality. Such assertions typically limit their attention to the material or organic aspects of personhood, and give no attention to the personal in interpersonal nature of human being. The fundamental flaw of this line of reasoning is to mistake the part for the whole and to reason to one's conclusions from a fragmented beginning. No such process can adequately capture the fullness of person, nor can it adequately answer the question of when life begins.

The End of Life

As with the beginning of life, attempts to define the moment when life ends, and when a person is no longer present are typically built on abstractions from the Personal, limitations of attention to the organic dimension of persons (which contains within it the material aspect but not the Personal). Historically, organic criteria of death have been used in medicine, previously the cessation of heartbeat and respiration currently, and, since the adoption of the Harvard criteria in 1968, the irreversible cessation of brain function. All of these definitions share of the common flaw of have attending to only one aspect of Persons, the biological, rather than Person as a whole and persons in

relation. Again, the part is mistaken for the whole and decisions are made based upon a fragmented vision of person.

Injured Lives: Neuropsychiatric and Neurological Illness

Finally, Neuroethics must continue to address issues about the care and treatment of individuals with neuropsychiatric illnesses, some of which present in a predominantly psychiatric arena (schizophrenia, bipolar disorder, psychotic disorders), as well as with various types of addiction, and some which present in predominantly neurologic ways, such as neurodegenerative diseases like Alzheimer's disease, Parkinson's disease, or in combined presentations such as those commonly seen in Huntington's disease and frontotemporal dementia.

For Neuroethics, the ethical demands of treatment of these individuals must move from a full accounting of personhood, to consider both personal and interpersonal factors, issues of capacity/competency, rationales for state intervention in psychiatric illness, and limitations of autonomy in individuals who pose a danger to society, as well as the use of medical resources, and the allocation of public funds for research.

VI. Conclusion: Neuroscience and Free Will in the Field of the Personal

In summary, Free Will exists, is an integral dimension of Personhood and can be recognized in persons fully conceived and adequately understood. In light of contemporary Personalist philosophy, the "problem" of Free Will is in the end one of our own making, brought about by efforts to abstract, that is, to limit attention about persons to one or more aspects that fail to adequately describe who a person is.

The "problem" of Free Will ceases to be a problem when persons are consistently and adequately apprehended in all their dimensions, the Material, the Organic, and the Personal, an apprehension that can occur when we move from a philosophy of thought to a philosophy of action. It also entails an examination of persons not in material or animal categories, but - following Burgos' lead - rather through categories unique to persons, including reason, freedom, and the capacity for self transcendence. When Neuroethics approaches persons in this manner, Free Will can be both recognized and preserved, Persons can be recognized as complex, integral and active, formed by and living in relation to others.

One implication of this process is that an adequate notion of person can only be attained through a multidisciplinary process in which science has an essential place but cannot be the only methodology; if it were, we would fall into the trap of abstraction from the Personal and approach decision-making in the domain of Neuroethics in a

fragmented fashion. By adopting a multidisciplinary approach that includes the physical sciences, biological sciences and the human sciences we guard against the trap of dualism, and against conceiving the world and ourselves in a fragmented fashion. Instead, we can develop an ethical vision in a Personal world in which freedom and determinism are not antinomies but instead dimensions of personhood. To do this is to fulfill Macmurray's criteria for an adequate philosophy, a vision of person that is logically coherent and adequate to the full range of human experience which is neither material nor organic, but a personal unity encompassing these dimensions.

A potential criticism of the position I have presented here is that locating the problem of free will in the Field of the Personal is simply transferring the same problem to a new place, leaving us with the same difficulties. While it is true that the question is here moved to a new context, that move does not mean asking the question in the same manner. By raising the question of Free Will in the Field of the Personal, and in a vision of person more complex than can be accounted for by a materialist worldview, it becomes possible to employ multiple methodologies (including but not limited to scientific reductionism, e.g. the social sciences, the humanities, legal studies) to provide a more complex, more nuanced vision of the activity of free will in human persons that more closely approximates the richness of that activity.

The purpose of this essay has been to consider the structural level of the free will debate, to examine the conceptual frameworks that have been employed to address free will and to ask if these structures (e.g., materialism, idealism, determinism, compatiblism, indeterminism) have been adequate to the task. My answer to that our conceptual architecture has been inadequate, leading to a situation of sustained conflict, because much of the contemporary debate has moved from a framework of physical and biological determinism that precludes a comprehensive discussion of the human capacity for free will. This framework has created a blind spot and a self-defeating conundrum; in order to argue for causal determinism, it is necessary to step outside material and biological determinism in order to argue that determinism is the whole story, thus creating a sustained logical contradiction. All proponents of determinism are persons, and can only make their arguments from the Field of the Personal, where free will resides, arguing in their activity as persons that such activity can not exist. At minimum, a broadening of our perspective can only serve to deepen the conversation and allow us to address substantive questions with greater depth and clarity.

References

Allen, Richard T. 2013. "Personalism and British Personalists," in *Three British Personalist Philosophers: Michael Polanyi, John Macmurray, Austin Farrer*. British Personalist Forum, www.britishpersonalistforum.org.uk.

Buford, T.O. 2011. "Personalism." *Internet Encyclopedia of Philosophy*. URL = http://www.iep.utm.edu/personal/.

Bengtsson, J.O. 2006. *The Worldview of Personalism: Origins and Early Development*, Oxford: Oxford University Press.

Burgos, J.M. 2013. *A New Personalistic Proposal: Modern Ontological Personalism (MOP)*. Paper presented at the 12th International Conference on Persons, Lund Sweden, August 7, 2013). (a copy of Burgos' paper is available for review; email jbeauregard@rivier.edu).

Burgos, J. M. 2012. *Introducion al personalismo*. Madrid: Ediciones Palabra.

Burgos, J. M. 2009. *Antropolgía: una guia para la existencia.* Madrid: Palabra.

Burgos, J.M. 2009. *Reconstruir la persona: Ensayos personalistas.* Madrid: Palabra.

Burgos, J.M. 2007. *Repensar la naturaleza humana.* Madrid: Ediciones Internacionales Universitarias.

Downie, Robin. 2002. "Personal and Impersonal Relationships." In D. Ferguson and N. Dower, eds., *John Macmurray: Critical Perspectives*. New York: Peter Lang Publishing.

Macmurray, J. 1992. *Reason and Emotion*. Amherst, NY: Humanity Books.

Macmurray, J. 1991a. *Persons in Relation*. Atlantic Highlands, NJ: Humanities Press International.

Macmurray, J. 1991b. *The Self as Agent.* New Jersey: Humanities Press.

Sardella, Ferdinando. 2103. *Modern Hindu Personalism: The History, Life and Thought of Bhaktisiddhānta Sarasvatī.* Oxford: Oxford University Press.

Williams, T. D. and Bengtsson, J. O. 2014. "Personalism." *The Stanford Encyclopedia of Philosophy,* E. N. Zalta (ed.), http://plato.stanford.edu/archives/spr2014/entries/personalism/.

APPENDIX

Major Trends in Contemporary Personalism

Personalism looks back to a broad range of thinkers in the history of philosophy, both in the West and in the East.[9] Focusing primarily on the Western tradition in this essay, Personalism in its contemporary sense has developed into number of distinct but related currents:

Communitarian personalism, grounded in the work of Emanuel Mounier, attends strongly to social action and social transformation.

Dialogical personalism, which gives emphasis to interpersonal relation as a ground for a philosophical anthropology, typified in the work of Martin Buber and Emmanuel Levinás.

American personalism, moving from an idealist philosophy and central European sources, building on the work of American philosophers Borden Parker Bowne, Edgar Sheffield Brightman, Peter Bertocci, and contemporary thinkers including Thomas O. Buford, Rufus Burrows, Randall Auxier, and European philosophers such as Jan Olof Bengtsson.

Hindu Personalism, emerging from Hindu philosophy and its search for freedom from suffering and addressing such questions as the nature of the self, touching on human agency, intention, free will and identity.[10]

British Personalism, focusing on philosophies of action typified in the works of Austin Farrer, John Macmurray and Michael Polanyi, and contemporary proponents such as Richard Allen and Charles Conti.

Islamic personalism, which shares some common roots with classical Greek philosophy, and examines both nature of the person and of God exemplified in the work of Muhammad Iqbal, Mohammad Aziz Lahbsabi and Allhagi Manta Drammeh.

---

9. Thomas O. Buford reviews the major philosophical strands in Personalism's background in his entry, "Personalism" in the Internet Encyclopedia of Philosophy, including its roots in the Greco-Roman period of the West, Hindu philosophy in the East, and has contemporary proponents writing in both of these traditions as well as Confucianism and Islamic philosophy. See, e.g. Gueye, C.M. Ed. 2011, *Ethical Personalism*. Frankfurt: Ontos Verlag.

10. See Buford, 2011, Section 1, "South and East Asian Personalism" for a description of Hindu and Buddhist tradition in India, China and Japan. Ferdinando Sardella has made a significant contribution to understandings of contemporary Hindu personalism in his recent *Modern Hindu Personalism: the History, Life and Thought of Bhakisiddhānta Sarasvatī* (Oxford: Oxford University Press), 2013.

Classical personalism looks to the Aristotelian-Thomistic tradition in the work of Jacques Maritain, Etienne Gilson, and contemporary philosophers such as Robert Spaemann and Thomas D. Williams.

Neopersonalism, which seeks and integration of classical and modern concepts of person in a new synthesis, found in the work of Czeslaw Bartnik, Luigi Stefanini, Maurice Nedoncelle, Edith Stein, Karol Wojtyla and Juan Manuel Burgos.[11]

11. The majority of these traditions are described by Juan Manuel Burgos in his recent summation of the (predominantly European) Personalist tradition in *Introducion al personalismo*.

Journal of Cognition and Neuroethics

Evolution Beyond Determinism: On Dennett's Compatibilism and the Too Timeless Free Will Debate

Maria Brincker
University of Massachusetts Boston

Biography

Maria Brincker is currently an Assistant Professor of Philosophy at University of Massachusetts Boston and has previously held an Arts & Neuroscience fellowship at Columbia University. Her interdisciplinary work spans a broad array of topics within philosophy of mind and neuroscience, such as affordance perception, mirror neuron theories, sensorimotor grounding of cognition, typical and atypical social cognitive development, action theory and aesthetics.

Acknowledgements

I want to thank the organizers and participants at the 2014 "Free Will" conference at the Center for Cognition and Neuroethics at the University of Michigan-Flint for their kind feedback on an earlier version of this paper, and for the Kane-Dennett Prize recognition and funding. The ideas of this paper have been stewing for more than fifteen years but this conference supplied the needed impetus to action. Further, I want to thank Elly Vintiadis for her helpful comments and last but not least deception expert Mark Mitton for our many heated discussions and his deep insights into questions of causality and the selective nature of perception. As Bergson, he has tirelessly alerted me to the impossibility of conveying the true dynamics of action with frozen symbols without hiding precisely what needed to be revealed.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Brincker, Maria. 2015. "Evolution Beyond Determinism: On Dennett's Compatibilism and the Too Timeless Free Will Debate." *Journal of Cognition and Neuroethics* 3 (1): 39–74.

Evolution Beyond Determinism: On Dennett's Compatibilism and the Too Timeless Free Will Debate

Maria Brincker

Abstract

Most of the free will debate operates under the assumption that classic determinism and indeterminism are the only metaphysical options available. Through an analysis of Dennett's view of free will as gradually evolving this article attempts to point to emergentist, interactivist and temporal metaphysical options, which have been left largely unexplored by contemporary theorists. Whereas, Dennett himself holds that "the kind of free will worth wanting" is compatible with classic determinism, I propose that his models of determinism fit poorly with his evolutionary theory and naturalist commitments. In particular, his so-called "intuition pumps" seem to rely on the assumption that reality will have a compositional bottom layer where appearance and reality coincide. I argue that instead of positing this and other "unexplained explainers" we should allow for the heretical possibility that there might not be any absolute bottom, smallest substances or universal laws, but relational interactions all the way down. Through the details of Dennett's own account of the importance of horizontal transmission in evolution and the causal efficacy of epistemically limited but complex layered "selves," it is argued that our autonomy is linked to the ability to affect reality by controlling appearances.

Keywords

Free will, determinism, compatibilism, naturalism, evolution, Dennett, causality, emergence, time, appearance-reality, process philosophy, action, choice, autonomy, self, metaphysics

Introduction: A tension in Dennett's compatibilism

Dan Dennett has over the past three decades developed and argued for an evolutionary and naturalist view of free will (Dennett 1984 & 2003). The core idea is that over evolutionary time, and through myriads of small and massively complex biological and cultural selections, the ability of voluntary control—what we call free will—has gradually evolved. Free will under Dennett's interpretation is simply the adaptive ability to anticipate outcomes and to flexibly exert control over factors in the world according to ones preferences, perceptions and deliberations. Beyond his discussion of how the ability of reason-based voluntary control has evolved, Dennett has spent considerable energy arguing that this kind of 'freedom' is 1) utterly compatible with determinism and 2) the kind worth wanting. He, in other words, defends a compatibilist position and works hard

to dispel the incompatibilists, be they hard determinists or libertarians, who all claim that "true" free will must involve some kind of indeterminism, and hence some kind of denial of determinism.

Dennett's view of free will might indeed be compatible with determinism. However, I shall in this article argue that his historical and perspectival evolutionary story of action choice and normativity would invite us to think beyond the classic determinism and indeterminism dichotomy, toward an emergentist or interactivist account. Further, I suggest that such an alternative would be more empirically plausible and, by way of it's temporality, makes room for a free will theory that even some indeterminists might find worth wanting.

Throughout his writings, Dennett employs a series of characteristic toy models or "intuitions pumps" as he calls them: "Intuition pumps are cunningly designed to focus the readers attention on "the important" features, and to deflect the reader from bogging down on hard to follow details" (Dennett 1984, 12). I propose that we might want to use a bit of Dennett's own methodology on himself and "turn the knobs" on his intuition pumps to see what is doing the work. Following his advice, I will "go slow where others go fast" and argue that his atomistic and compositional metaphors might be doing a lot of the heavy lifting in his compatibilist picture. I point to tensions within Dennett's own account between his keen eye to the causal role of layered heterogeneous inners, as well as the causal efficacy of pattern recognition in evolutionary and cultural history on the one hand, and his eternalist and compositional atomist approach to questions of metaphysics on the other. The question is whether without his metaphysical assumptions, his evolutionary story would invite a more temporal, emergentist—and arguably more naturalist—understanding of the causal fabric of the world. Thus, his own evolutionary theory might be used to reveal new options in the metaphysical possibility space, and thereby show us beyond the deadlocked dichotomy of classic determinism and indeterminism.

A. Dennett's naturalism & critique of "the buck stops here" claims

Dennett is a vehement critic of libertarian free will positions, as these typically point to some arbitrary break in the causal chain. He does not exclude the possibility of indeterminate quantum events nor that some level of indeterminacy or randomized processes might be useful—and thus evolutionarily, culturally and individually selected for—but he argues that no *actual* indeterminacy is needed at any relevant level of description to explain action choice. In other words, Dennett argues that indeterminacy

can be seen as either an irrelevant feature of ultimate physics or as a tool within a determinist story. He is particularly suspicious of the notion of "agent causation," and the idea of any agent-level claim of indeterminacy. He argues that such agent-as-unmoved-mover, "Deus ex Machina," "the buck stops here" claims neither make any scientific sense nor—as indeterminism is precisely uncontrolled—seem to give us the kind of "free will worth wanting" (Dennett 1984 & 2003). Dennett has similar objections to other "the buck stops here" claims, and e.g. in connection with his analysis of intentionality, he compares the idea of "unmoved mover" to "unmeant meaners" and irreducible "original intentionality" (Dennett 1989, 288). He sees such claims as a sort of unscientific mysticism or at least as arbitrary and empirically intractable. His project is to show that we can make sense of our experience of action control, lived possibility and responsibility without "magic feathers" just like he thinks that consciousness can be understood without reference to a "Cartesian theater"(Dennett 1991).

In spite of his claims to the contrary, many in the conceptual landscape of the free will debate view Dennett's claim that our free will is compatible with determinism as incoherent and eventually interpretable as an instance of hard determinism and in effect implying that we do not as a matter of fact have true free will.[1] As we shall see, a central unsettled issue is the question of whether, and in which sense we are free to "do other wise." Dennett says that even if we cannot do otherwise from some sort of metaphysical Gods-eyes perspective, we have real lived options from our own epistemically limited perspective. Critics suggest that this makes free will an illusion, but Dennett insists that free will depends not on metaphysics but on our lived choices. He argues that our ability to perceive and deliberate about competing action possibilities is integral to action, anticipation and control, and these abilities all depend on eons of gradual processes of biological and cultural selection and horizontal transmission. Thus, our processes of decision-making should neither be seen as illusionary nor as causally epiphenomenal according to Dennett, but rather as an integral part of the causal fabric of everything. Dennett even challenges the view that we should care if we could have really done otherwise under the particular cosmic microstructure. He argues that the presence or absence of metaphysical possibilities would be utterly inscrutable from our lived perspective, and therefore should not cause us any worries. "The kind of free will worth wanting" is not metaphysical but practical. The key is our biological, cultural

1. See e.g. his recent review of Sam Harris' book see here: http://www.naturalism.org/Dennett_reflections_on_Harris%27s_Free_Will.pdf and for Harris' reply: http://www.samharris.org/blog/item/the-marionettes-lament.

and political ability and freedom to act according to our own—equally biological and cultural—preferences. Dennett thus uses, the Gods-eye point of view to argue for the compatibility of determinism and free will, but challenges the psychological relevance—not consistency—of the metaphysical view, which makes determinism inacceptable.

2. The gradual evolution of freedom

Before further discussing the metaphysical aspects of Dennett's compatibilism, we need to look at Dennett's empirical and evolutionary account of the historical complexity underlying processes of human action control. I think that the millennia-old free will debate is in dire need of some empirical details about how we actually do seem to choose and control our voluntary actions. Note though—Dennett is not simply interested in revealing the details of the evolutionary histories, but is also concerned with using these as evidence to show that the whole can come to be "freer than its parts"—and by implication that freedom could have emerged in a determinist world.

> It seems to stand to reason that nothing composed of such unfree parts could have any more freedom, that *the whole cannot be freer than its parts*, but this hunch, which is the very backbone of resistance to determinism, will turn out, on closer inspection, to be an illusion. (Dennett 2003, 61)

Though we might not agree on the nature of the "parts," I share Dennett's belief that freedom has and can evolve—and that in that sense that the whole can be freer than its precursors. I also agree with his critique of essentialist and unchanging categories. The world is filled with borderline cases of everything, and, as Dennett, I see this variance as a crucial feature of our world rather than something that needs to be explained away.

A. Historical precursors and relational categorizations

In *Freedom Evolves* Dennett discusses the chicken-egg paradox of the "first mammal" and the idea that—even though the parent of a mammal must per definition be a mammal—going back in evolutionary time, at some point our 'parents' were not mammals. Does that undermine the fact that we are mammals? He writes:

> What should we do? We should quell our desire to draw lines. We don't need to draw lines. We can live with the quite unchocking and unmysterious fact that, you see, there were all these gradual changes

> that accumulated over many millions of years and eventually produced undeniable mammals. (Dennett 2003, 127)

Note Dennett's pragmatic terminology: We "can live with" gradual changes and yet at some point the category is "undeniable." The category, thus, emerges gradually and its eventual factuality is based in the relational response—not an essentialism of inherently necessary and sufficient properties. Dennett's advice to us is to "quell our desire to draw lines." I agree. Drawing neat lines is a powerful epistemological activity that often yields problematic metaphysical inferences. However, I shall suggest that 'line-drawing' i.e. discrimination and discernability, might be the engine of causality and change. As we shall see, I am concerned that Dennett's own "intuition pumps"—by insisting on atomistic/digital (all-nothing) and compositional (part-whole) building blocs at the lowest level of reality—might lose the baby with the bath water. That is, he might forget his own insight that category borders typically are products of line-drawing activities, which themselves have to be explained. More on that later, but his critique of essentialist and unchanging all-or-none categories is well taken.

B. The evolution of normative entities and interests

Dennett's project is to show that freedom can evolve—and not freedom in some sense of random undetermined action—but rather freedom as the ability to act according to ones interests and preferences. But what are *interests* and what does it mean to say *someone* has interests? Immediately we have opened two core cans of worms: 1) the issue of normativity/teleology and 2) the issue of what it means to be a self. Here again, we see Dennett's anti-essentialism and his historical approach. The precursors for what Dennett purposively anthropomorphically calls "interests" did not evolve over night but slowly emerge through complex processes. He writes:

> The day the universe contained entities that could take some rudimentary steps towards defending their own interests was the day interests were born. The very tendencies of these organisms to preserve this or that (their varieties of *homeostasis*) helped sharpen the definition of their interests. Only certain sorts of homeostasis tended to be self-preserving in the long run; those kinds were replicated hence persisted, and hence gave further definition to the crude primordial "interest" in self-preservation and self-replication. (Dennett 1984, 22)

Like in the question of the primordial mammal, we see how Dennett wants to give a gradual multipronged story of a complex generative dynamic process of evolution. The circularity and self-referential nature of the explanation is also apparent: What does it mean to be a rudimentary *self*? To be an entity that takes steps towards *self*-preservation and *self*-replication. What does it mean to have rudimentary *interests*? To be an entity that acts to protect (preserve and replicate) those *interests*. What breaks the unfruitful circularity of teleological explanation is its fruitfulness: I.e. the hindsight of preservation and replication. Judged from the present, we can look back and reify gradual "interests" and "selves" as precursors just like we can indentify non-mammalian mammalian ancestors. Note however, the different meanings of self-preservation given the inclusion or exclusion of self-replication. What is it that "persists" and that is preserved? In the case of a single organism, it is the survival of the organism and in the case of persistence through replication, it is the persistence of the organization and historical line of this organization through it's expression in future individual organisms. Is it a bit more of a stretch to talk about replication-based persistence in terms of *interest* and *self*-preservation? Dennett does not seem to think so, as he is rather comfortable with what he calls the "gene-eye-perspective" (and later "meme-eye-perspective"). As a matter of fact, he often takes this non-phenotypical "perspective" at sub-personal levels. But what would it mean to say that genes are selves with interests? Dennett himself points out the difference between us having interest in, and taking steps towards, preserving say a rock formation and the notion that this formation had interests of its own (1984, 22). Thus, doesn't it even go against Dennett's own stated definition to think of genes as carriers of interests, as genes are non-autonomous and thus incapable of taking any, even ever so rudimentary, "steps towards defending their own interests"?

Dennett is well aware that a 'naked gene' is not an agent—not even in the primitive precursor sense. This is also clear in his discussion of memes. Dennett follows Dawkins and draws an analogy between biological genes and cultural "memes" as replicating recipes for behavior—or as he writes "a meme is an information-packet with attitude." Language structures are examples of memes, and speech thus a 'memetically'-based human cultural activity, transferred between people and generations. But like the "naked genes" of viroids, Dennett stresses that they depended on the human organism to be "expressed."

> Memes depend on human brains as their nesting places: human kidneys or lungs wouldn't do as alternative sites, because memes depend on the thinking powers of their hosts. Being involved in thinking is a memes

> way of being put through the paces and tested by natural selection, just as getting one's protein recipe followed and getting the result out in the world is a gene's way of being tested. If memes are tools for thinking (and many of the best of them are just that), they still have to be wielded for their phenotypic effects to show up. You still have to think. (Dennett 2003, 186)

As we see in this passage, Dennett stresses the dependency of meme and genes on organisms for their phenotypical expression and thus evolutionary "selection," and further he highlights their nature as "tools." But he simultaneously takes what he calls the "gene-eye" and "meme-eye perspective" by writing "getting *one's* recipe followed." One might here want to ask what these purposively anthropomorphic "intuition pumps" are doing for Dennett's argument. Firstly, it seems that he wants to use these analogies to problematize the uniqueness of the human agentic perspective. He wants to give a story of gradual evolution of capacities rather than a radical all-or-none appearance of the human intellect. This part of the project I fully support. However, the question is if Dennett is so focused on the notion of "replication" that he downplays the dynamic processes of the autonomous "self" that is to be preserved or perhaps replicated? Dennett is perhaps using the gene-eye point of view as a way to resisting what he sees as mythical and Cartesian notions of a too substantial personal level self. The gene and the meme are "entities" that can be studied, and might be seen as the "things" that are being replicated. Dennett might use these as fix-points in his attempt to show that personal level teleology, preference and selection has biological precursors, and that "the freedom of the whole can be freer than the parts." In other words, he appears to be using the "gene-eye perspective" to imply that the world, minds and actions can be understood via a sort of atomistic compositionality. Ironically, I shall argue Dennett's seeking of anchoring "things" and "parts" brings in a new set of "unmoved movers" and might obscure rather than help our understanding of emergent "selves," autonomy, teleology and subjectivity. But to make this point, we need to look deeper into the evolutionary story.

C. Horizontal transmission & emergence of Bastardized 'inners'

In Dennett's attempt to give a bottom-up story, he does not just look at the last efficient cause of an action but seeks to understand freedom through the messy web of actual causal histories. A part of this deep history are the big leaps of evolution through horizontal integrations of different lineages, of "self" and parasitic "non-self" entities. Dennett here again draws parallels between biological and cultural horizontal

integrations.[2] I share Dennett's view that there is a continuity between the biological and cultural domain, and further that the details of the various layerings via horizontal integration are crucial to the multiplicity of timescales of regulation and control of multicellular organisms. I shall focus on his analysis of how the evolution of eukaryotic cells though the invasion and survival of one prokaryotic cell inside of another have allowed for new orders of complexity (2003, 144–146). This evolutionary history of layering is worth dwelling on as it might help us understand some basic dynamic characteristics of self-regulating organisms. One such question that Dennett discusses is how prokaryotic cells somehow already came with much more complex gene-regulation processes than they needed. He writes:

> Fancier eukaryotic cells, however, to say nothing of us multicellular types composed of these more complex building blocks, need a mind-boggling elaborate system of intermediate steps, checks and balances, so that genes can be turned on and off at appropriate times by the indirect effects of other gene products and so forth. For some time biologists had a classic chicken-and-egg puzzle to contend with: How did this elaborate gene-regulation machinery evolve? Multi-cellular life couldn't even begin to evolve until most of this expensive machinery was in place, but it apparently isn't required for prokaryotic life. (Dennett 2003, 146)

In other words, how could these much needed cellular control mechanisms be in place prior to the first eukaryotic pre-cursor cells? Such control mechanisms, balancing and adaptively applying tools given changing circumstances, allow primitive multi-cellular organisms some level of autonomy and ability to have responses selected from various repertoires. Thus, their evolution might give us an important clue also to the understanding of our human-level abilities of action control. Dennett explains the emergence of this sort of cellular complexity in prokaryotic cells as follows:

> The answer that is now emerging is that it was paid for by a civil war that raged for roughly a billion years of early prokaryotic life. It was an arms race within the genome, with good citizen genes doing battle with those transposons – free loaders who copied themselves repeatedly in the genome without providing any benefit to the whole

2. He writes e.g. "Horizontal transmission of design, of information that can be put to good uses, is a key feature of human culture, and undoubtedly the secret to our success as a species" (Dennett 2003, 145).

> organism. This created lots of measures and countermeasures, such as silencing mechanisms and isolation-defeating mechanisms. (Dennett 2003, 146)

This is indeed an interesting history to understand. Note the idea that the large jumps in evolutionary novelty became possible due to active processes of stabilizing the cells in the face of inner turmoil at an earlier stage. Further note, that when brought under control this intruder-created nuisance produced new potential capacities, which would then later be exploited to the organism's advantage in entirely new contexts. A case of what doesn't kill you makes you stronger—here in the specific sense of having expanded the repertoire of the 'self'—and thus, response possibilities. Such inner complexity is expensive as Dennett says, but also in a broader sense opens you up to new kinds of vulnerabilities. I therefore like to call these events "bastardizations," implying at once the multiple origins, otherness, and the always-looming possibility of indigestion, exclusion, isolation, general friction and precarious reception. But it also implies transformation, creative novelty and the birth of a truly new and unique self, the main challenge to whom comes from the quest of social recognition i.e. fitting existing categories. It is thus always a fragile balance of how much can be metabolized without collapsing the autonomy organism.[3] Ironically, our subjectivity and great evolutionary assent to freedom has come through a series of debasing and tumultuous internalizations of others.[4] But this is of course me projecting my interpretation onto Dennett, as his metaphors are generally more about wars, infections and peaceful co-existence than about illegitimate children and the precariousness of subjectivity. He writes:

> The eukaryotic revolution draws attention to the fact that even in biological evolution, which Darwin aptly called "descent with

3. One might also suggest that the emergence of inner turmoil in moving organisms might have presented a puzzle of how to create coherence and negotiate the now necessary frame-shifts between inner and outer concerns. A puzzle to which conscious feelings and perceptions might have been an evolutionary solution.

4. We are at present exposed to a radical cultural horizontal transmission through the technological inundation of information and surveillance. These things give us powerful new action repertoires and reach but also transform us—and our worlds—at runaway speeds. The borders of 'selves' and 'societies' are changing in ways that have significant effects on our autonomy, and if we don't evolve new 'membranes' to help us control, regulate and hide our inner multiplicities, then we might very well end up adapting by moving the prime locus of autonomy to a societal level. In other words, we are in a period of externalization of our inners, which as we shall see might represent a loss of freedom in the sense of a loss of the ability to hide.

> modification," there is plenty of room for horizontal transmission of design. The prokaryotic hosts who were first "infected" by their symbiotic visitors got a huge gift of competence designed elsewhere. That is, they didn't get all their competence by vertical descent from their ancestors via their parent and grandparents and so forth. They didn't get all their competence from their genes, in other words. They did however pass on this gift to all their offspring and grand-off spring through their genes, since the genes of the invaders came to share the fate of the nuclear genes of their hosts, traveling side by side into the next generation, which was infected at birth, one might say, with its own complement of symbionts. The clear trace of this dual path is still highly salient today, in all multicellular creatures, including us. Mitochondria, the tiny organelles that transform energy in each of our cells, are the descendents of such symbiont invaders, and have their own DNA. Your mitochondrial DNA, which you get only from your mother, exists in each of your cells, alongside your nuclear DNA – your genome. (Dennett 2003, 145)

Given this passage, the crucial question is when something becomes part of "me." Dennett's terminology here suggests that 'I' am my genetic code but that 'I' depend on these ancient invaders, which are replicated in my offspring by an "infection at birth" but stay forever foreigners traveling "side by side" as *their* DNA has never merged with *mine*. But isn't an alternative interpretation of the same biology, that I exist as a genetic multiplicity, that *my* genomes are a combination of both the mitochondrial and the nuclear DNA? After all, even the nuclear DNA is a bastardization as mentioned above. Interestingly Dennett leaves out the story of what one might consider life's 'original bastardization'. One candidate could here be the first trapping of RNA (possibly itself imported via meteorites from Mars[5]) in primitive lipid vesicles, and thus perhaps a primordial self of *self*-replicating life. However, the question is if Dennett would be more prone to think of the earlier RNA duplications as already self-replicating. The difference is that only with the event of the proto-membrane can we talk about some sort of inner environment, and also of clear processes of self-maintenance and autopoesis. I shall not suggest that we can discern a primordial mammal or a truly original bastardization of life, but merely highlight that Dennett's evolutionary accounts have a tendency to downplay

5. See Grossman & Webb (2013).

or sideline the dynamics of boundaries and habitats in the evolution of self-replicating individuals. He does in the quote above talk about the necessity of the wider "molecular machinery," in bringing about the phenotype. But yet he talks as if the gene/meme is the individual or the 'entity' that is being (self)-replicated, and thus, as somehow the locus of "interests" and somehow a "self." In brief, he spends a lot of time on genes and memes as recipes for self-replication, but comparatively little on their contexts. References to membranes, time and autopoesis are suspiciously missing from the index of his books—as if he tells the story of the chicken but forgets to mention the egg.

These terminological choices are of some consequence when considering the issue of what it means to be an agentive self— and whether the DNA is more important than the boundary that actively establishes and maintains a difference between "self" and "other" not by any inherent pre-given essence or difference but via on-going dynamic and non-instantaneous interchanges. With the gene as the locus of the proto-self self-replication, Dennett avails himself of a relative temporal stability and passive self-containment. If we on the other hand focus to the entire cell or organism we can see that its stability, maintenance and autonomy is a precarious, active and contextual process in constant flux.

D. The spatio-temporal agent is the loop

As we have seen already, Dennett wants to dispel the need for extra supernatural "magic feathers." He sees our abilities to recognize, anticipate and select actions as tools to maintain some preferred homeostasis or outcomes that we have developed and gradually selectively shaped through both evolutionary time and though our cultural environments. Based on analyses like the above one of the eukaryotic revolution, Dennett defends the idea that our human-level freedom can be understood analogously, as much fancier ways of using meme invader and "trapping reasons," for purposes of intricate forms of anticipation and avoidance. He argues that all this is possible to understand without any super-natural assumptions or breaks in the causal fabric of reality; "the whole can be freer than its parts," and we might add *the present more free than it's past*:

> Events in the *distant* past were indeed not "up to me," but my choice now to Go or Stay is up to me because its "parents" – some events in the *recent* past, such as the choices I have recently made – were up to me (because *their* "parents" were up to me), and so on, not to infinity, but far enough back to give my self enough spread in space and time so that there is a me for my decisions to be up to! The reality of the moral

> me is no more put in doubt by the incompatibilist argument than is the reality of mammals. (Dennett 2003, 135–136)

We see here, the analogy between this notion of autonomy of something being "up to me" and Dennett's primordial mammal argument. The purpose of the analogy might be to pump the intuition that autonomy too is not only a category with degrees—an intuition I share—but also that it is a historical lineage category. I also agree that autonomy is historical, but like we saw in the discussion of the gene eye-perspective, I think we are genetic bastards and that autonomy is continuously re-negotiated in the present. Thus, I am hesitant to endorse the idea that whether my current action is up to me depends that strongly on whether its precursors has the proper kind of history.[6] However, Dennett also adds that the question of whether we see something as being up to us depends on whether we go "far enough back to give my self enough spread in space and time so that there is a me for my decisions to be up to!" This point about how big or small a perspective we take on our selves is really fascinating and something that Dennett returns to multiple times in discussing moral responsibility.

The general idea of a both spatially and temporally extended self also plays a key role in Dennett's argument against the incompatibilists, as it is via this notion that we become the authors of our actions and that we play a consequential causal role in the world. Dennett argues that the inference, that our heart-wrenching deliberations, thoughts and actions would be obsolete in a determinist world, comes from the idea of an extensionless and non-material self. In accordance with his critique of the Cartesian theater, Dennett is similarly critical of the idea of some sort of localizable atomistic conscious self, which can be neatly contrasted with the non-self aspects of our physiology. Hence, in his discussion of Libet's findings and assumptions about the causal role of the conscious self he concludes: "What Libet discovered was not that consciousness lags ominously behind unconscious decision, but that conscious decision-making takes time" (2003, 239).[7] And a bit further down he writes:

> When we remove the Cartesian bottleneck, and with it the commitment to the ideal of the mythic time t, the instant when the conscious decision happens, Libet's discovery of a 100-millisecond veto

6. Note the similarity to Millikan's teleosemantics and historical notion of "proper function"—and perhaps the challenges her account faces due to its disregard for the influence of synchronic organization. (Millikan 1989).

7. Gallagher (2006) has made similar criticisms of the implausible temporal assumptions of the Libet studies.

> window evaporates. Then you can see that our free will, like all our other mental powers, has to be smeared out over time, not measured at instants. (Dennett 2003, 239)

I couldn't agree more, both with the claim that the notion of a "time t" for action is a myth, and the claim that free will has to be smeared out over time. In the further explanation, we can really see the central role the extended self plays in Dennett's compatibility theory both our freedom and responsibility. He writes:

> Once you distribute the homunculus (in this case, decision making, clock watching, and decision-simultaneity-judging) in both time and space in the brain, you have to distribute the moral agency around as well. You are not out of the loop; you are the loop. You are large. You are not an expansionless point. What you do and what you are incorporates all these things that happen and is not something separate from them" (Dennett 2003, 241–242)

This is an extremely important idea—if we are not Cartesian dualists or in other ways thinking of free will as external to the material world and the biological body, then, in so far as "we" exist, it is natural to think of our deliberations and decision-making as not only being *in* the causal loop but *being* the loop. It is thus by way of this idea that Dennett claims that his account of voluntary action indeed makes actions, not only up to us, but causally dependent on us. Further, he writes, "you have to spread the moral agency around as well." Thus, the degree to which we see ourselves as large or small corresponds to our respective internalization and externalization of the responsibility. His point seems to be mainly a question of perspective: if we on the one hand see ourselves from the outside we tend to look at the multitudes of "external" causes of our action. But if on the other hand we spread ourselves out, then those same causes seem to "internal" to ourselves and they would thus be our responsibility. The question is whether this leads to an unnecessary moral relativism. Are all perspectives equally morally valid? In other words, does Dennett in his eagerness to avoid the homunculus again forget about the semi-permeable membranes and the phenotypes selection works on?

E. Appearance and reality—the causal effects of hiding and revealing

I want now to highlight another dynamic—and possibly constitutive—aspect of autonomic evolved selves. Namely, the ability to selectively hide and reveal aspects of both our inner and outer worlds and thus purposively control interactions. Dennett

himself talks about intentional deception as an important mental capacity as well as of course our ability to track and gage when we and others are being deceived. He proposes that the difference between appearance and reality is "fatal" to all organisms, but only we humans have the ability to reflect on and deliberately "bridge the gap."[8] Dennett is probably right that our ability to voluntarily take steps to make reality appear and not appear sets us apart. However, he is also well aware that many unconscious steps towards controlling the gap between appearance and reality are at the core of the evolution of life. He writes e.g.: "Mother nature abides by the "Need to Know" principle"(1984, 24). Thus the issue is not simply the human ability to deliberately "bridge" —i.e. eradicate— the gap but rather to purposively control and exploit the gap.[9]

Further, I would argue that one could see this control as a core ingredient in the emergence of teleological evaluative selves and thus as a key to any evolutionary story of the freedom of action. The question is if it isn't the "hiding" of the genetic material behind a membrane that allows for steps towards self-preservation? If the world simply triggers the replication is it self-replication? Independently of how one decides to categorize the primordial mammal or the beginning of self-replicating life, I think one gets the picture wrong if evolving processes of control are not seen as dependent on and contributive to abilities of hiding and revealing.

Dennett's story of horizontal transmission in both biology and culture is fascinating to see against this background—as we then see that the self-other dynamic is one of control. What is "me" is not my just my nuclear genome and it is not the conglomerate of my human stem-line cells with both mitochondrial and nuclear DNA, but rather a much more genetically and culturally messy—and constantly changing—set of processes. Are the symbiotic microbes of my digestive system part of me? Is my foot? Is my language?

8. "The difference between how things really are is just as fatal as a gap for them as it can be for us, but they are largely oblivious to it. The recognition of the difference between appearance and reality is a human discovery. A few other species—some primates, some cetaceans, maybe even some birds—show signs of appreciating the phenomenon of "false belief"—*getting it wrong*. They exhibit sensitivity to the errors of others, and perhaps even some sensitivity to their own errors as errors, but they lack the capacity for the reflection required to dwell on this possibility, and so they cannot use this sensitivity in the deliberate design of repairs or improvements of their seeking gear and hiding gear. That sort of bridging of the gap between appearance and reality is a wrinkle that we human beings alone have mastered" (Dennett 2003, 165). Note that Dennett's thoughts on human and primate abilities to appreciate false belief have lead to a whole industry of test paradigms. See also Brincker (2014) for a recent theory of the development of "false belief" implicit and explicit performance in humans.

9. See also Dennett points about self-deception (1984, 48).

Are my contacts? The point is not to answer these questions in the abstract. The point is that our bodily reactions, engagements and actions shape the answers as they answer. I am always a self—or rather selves—in the making. My complexities and messy causal histories determine me and yet scaffold my freedom.

F. Options & evitability

Dennett thus, builds forward to his notion of freedom as the ability to recognize, anticipate outcomes and actually choose what you prefer—given your also evolved preferences.

This process, he argues, can be seen as entirely deterministic: You recognize what he calls "a "special interest opportunity" and given your anticipation of outcome compared to your preferences you select either to act towards or to avoid. He writes

> So a real opportunity is an occasion where a self-controller "faces"—is informed about—a situation in which the outcome of its subsequent "deliberation" will be a decisive (as we say) factor. In such a situation more than one alternative is "possible" so far as the agent or self-controller is concerned; that is, the critical nexus passes through its deliberation. (Dennett 1984, 118)

Thus, Dennett takes possibilities for action to be recognized opportunities that one could either decide to pursue or not. Accordingly, he argues that determinism does NOT, as most claim, make actions "inevitable." Rather, we have evolved—possibly in a determinist fashion—to be very fancy evators or avoiders.

Both hard determinists and Libertarians here typically object and say that this is not true "evitability" as you "could not have done otherwise"! Dennett says no not in *these exact circumstances,* given these precise preferences, perceptions and anticipations this precise action was selected by your "will" if you want. Had it been a different—even ever a so slightly different—universe you could have done otherwise. And Dennett says that we always sneak in such different circumstances when we say we could have acted differently. Thus e.g. *yes had I wanted to* I could have picked a different action…but his point is that this different volition would exactly be a product of a different causal world—not this world.

I will in the following sections analyze Dennett's metaphysical arguments and the intuition pumps he uses to defend this idea that other outcomes could only come about in other possible worlds. A closer look seems to reveal a tension within Dennett's compatibilism. As we have seen in his evolutionary story he points to 1) the historicity

and precursors of categorizations, and the causal efficacy of 2) our "loop" of perspectival epistemic limitations, spread-out layered selves and 3) of the gap between appearance and reality. Yet, he assumes a metaphysical base level of non-perspectival causality where appearance and reality always neatly coincides. I shall suggest that he overlooks a set of metaphysically possible worlds in which all causal interactions are limited and there thus—even "under the aspect of eternity"—is no base-level, but always precursors and "back stages" to use Dennett's own term. But I get ahead of myself.

3. Dennett's metaphysical view from eternity

Dennett argues that from our lived perspective we have multiple possibilities for action, but these are not possibilities to actually change the course of the world. Quoting from *Elbow Room*:

> But if we want to change the course of history we are in for a big disappointment. For no one can change the course of history – for reasons that have nothing to do with determinism. At the beginning of the chapter we imagined all of space and time, past, present and future laid out before us ("*sub specie aeternitatis"* in philosophical parlance: under the aspect of eternity). If the scene we thereby imagine is supposed to be the *actual* course of history through eternity, then – look, and see – the image has no branchings. Only one actual thing happened whether or not what happens is determined to happen, so the part of our image we label "Future" consists of the events that actually happen – happen to happen – in the fullness of time. (Dennett 1984, 124)

Dennett is right that branches are hard to make sense of—as Bergson (1889) pointed out their imagery assumes the hindsight of all the alternative actions as actually having been carried out. However, many would say that the kind of free will worth wanting is precisely one whereby my actions can change or add to the course of world history. Dennett would likely respond that my actions do participate in the causal history of the world and thus influence it, but "under the aspect of eternity" I have changed nothing. I argue that this view from eternity is a fancy vending machine.

A. The metaphysical stance

I have admittedly always been struck by the oddness of the age-old "metaphysical stance," above conceived of as "imagining all of space and time" "laid out before us." It is supposed to be the ultimate exercise of objectivity of abstracting away our situated perspective—and yet it seems impossible to formulate this metaphysical notion in abstraction from some sort of perceiver or subjective point of view. This feature of an entirely external non-specified perceiver is clearly seen in the expression, "the view from nowhere" and the "God's-eye perspective"—or here as Dennett writes "under the aspect of eternity." The purpose of the non-specificity of the perceiver is presumably that any specification would impose a limitation to the view and thus prevent the access to the totality of being, the "Ding an sich." It is thus a paradoxical mind-bending activity of imaging the mind-independent world: the appearance of reality independently of appearance, i.e. the ultimate bridging of the gap between appearance and reality.[10] But what could it mean to "see" (not to mention smell), the entire "actual course of history"? As Akins (1996) reminds us, our senses are narcissistic, and it seems that the metaphysical stance invites us to abstract from *any kind of categorization*, *recognition or reification process* that we would rely on in actuality. This "metaphysical stance" exercise is at least as old as the ancient Greeks and likely as old as metaphysical thinking, and is constantly appealed to in the free will debate as a way of exposing the inconsistency of a determinism and the notion of free will as the ability to do otherwise or "agent causality."

It should be stressed that it is a pivotal part of Dennett's own compatibilist argument that we can *never actually* reach a God's eye point of view—and that this epistemological limitation is precisely what makes our action-opportunities real from our perspective. Laplace's omniscient intellect is impossible in reality and this makes a difference to Dennett, as it guarantees the practical unpredictability of the world. Even if the world is entirely determinist, as long as it is inscrutably so and forever hidden, it is irrelevant to our freedom.

Yet, Dennett appeals to this admittedly impossible vantage point to deny the very possibility of metaphysical possibility. So, we need to understand how this image is doing

10. See also Putnam's critique of what he calls the "externalist view." He argues that there cannot be "exactly one true and complete description of the 'the way the world is'…there can be no God's eye view of reality" and writes further: "What we have here is the demise of a theory that lasted for over two thousand years. That it persisted so long and in so many forms in spite of the internal contradictions and obscurities which were present from the beginning testifies to the naturalness and the strength of the desire for a God's eye view"(Putnam 1981, 74). See also Hornsby on the nomological character of causality (Hornsby 1997, 78–80).

work in his argument. I will therefore look closely at his metaphysical images and models, which are used 1) to evoke the idea that the future is closed, that we can change nothing and 2) to explain the determinism, which he claims to compatible be with lived freedom.

B. Laplace's infinite intellect

Before we turn to Dennett's use of the metaphysical stance, I want to say a bit about Laplace's famous "metaphysical stance" image of determinism. He writes: "We may regard the present state of the universe as the effect of its past and the cause of its future." And to explain this determinist causation from "the present state" to the "it's future," he continues:

> An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes. (Laplace 1814)

Thus, the way Laplace conceives or "sees" the past as fixing the present and future is through an exhaustive knowledge of a time-slice of a frozen universe set in predictable motion. Note that the role of the "intellect" is as the measuring stick, the background (or foreground) in relation to which—before which—the world is arrested and everything is reified. In other words, the picture does not work without its implication of an intellect or some kind of "reifyer" which—as any measuring stick—itself is external to the universe, not included, not itself "given." It is the full transparency (appearance=reality) of everything *for* the intellect that allows for the exhaustiveness of the knowledge and thus predictive completeness. With limited and perspectival transparency uncertainty would not be eradicated and the world would appear only partially and thus not be transparently "present" as a totality. Interestingly, Laplace's image imposes a temporal limitation to the otherwise vast intellect. The intellect need not know everything "under the aspect of eternity" but only know everything at "a certain moment," and then given the deterministic analysis of forces and motions the temporal transparency appears: everything becomes "present before its eyes."

The notion of time slices, moments, instants, as we shall see in Dennett's analysis, seem to be a fixture of determinism formulations. Laplace takes the instantaneous transparency of everything for granted. But it is an open question if all causalities

and regularities of "the greatest bodies of the universe and those of the tiniest atom" could possibly be arrested and "revealed"—even to an infinite intelligence—on one homogeneous and near instantaneous temporal scale. This is important, as the purpose of image is to argue that *the actual world* we live in—behind the appearances—is fully determined and timeless in the sense that it could have been predicted and instantaneously "present before" us if we knew the initial conditions of the universe.

C. The view from eternity & the denial of possibilities

As we have seen, Dennett proposes a historical, empirically-based theory of how our ability of voluntary action has evolved and how it can be understood as arising within a causal understanding of nature. Free will, he argues, does not break the causal web as most have assumed, but rather acts within it. However, note that his metaphysical view of the world does not depend on determinism. He writes: "The image we should have when we imagine this universe to be nondeterministic ought to be indistinguishable from the image we imagine when we conjure up a deterministic universe through time" (Dennett 1984, 124). How so? The key to Dennett's suggestion, that the future's closedness does not depend on determinism, seems to lie in his eternalist "bloc universe" view. Though a defender of determinism as compatible with free will, he is agnostic about whether the world from an inaccessible point of view unfolds in a determined or indetermined fashion. Thus, the reason why he sees the metaphysical stance as excluding changes to the course of history is *not* that we are caught in a determinist causal web, in which all actions in theory could have been predicted from initial conditions as Laplace's famous thought experiment would have it. He continues:

> If the past is unchangeable, the future is unavoidable – on anyone's account. The future consists, *timelessly*, of the sequence of events that will happen, whether determined to happen or not, and it makes no more sense to speak of avoiding those events as it does to speak of avoiding the events that have already happened. (Dennett 1984, 124)

Under the eternalist view, there is no passage of time, just a totality of moments or temporal extension. Time is a dimension like space, and therefore the past and future are equally fixed and "timelessly" existing. Dennett asks us to "look" at the totality of time: "past, present and future laid out before us." Note that Dennett here is assuming a different extent of transparency (appearance=reality) than Laplace did. He immediately jumps to the view from eternity, which Laplace only gained through the application of deterministic analysis. From that vantage point, causes and laws are irrelevant and

the indeterminist and the determinist universe are equally transparent. So, what has Dennett shown with his eternalist view? That the passage of time changes nothing? That the "present" holds no possibilities, nothing more than what "will happen"? It seems to me though, that these conclusions are assumed in the formulation of the timeless eternalist picture. In other words, isn't the claim that there could be local possibilities, which can cease to exist, precisely denying that one can see everything from nowhere, from eternity, from God's house, you name it? If one assumes that *everything* is "seen" from this metaphysical stance then one limits the possible metaphysical options accordingly. Only worlds of full eternalist transparency are possible. The transparent moments could be "sequenced" in an either determinist or indeterminist fashion, but there shall be nothing hidden, no time, no change of what has already been assumed. Any view that takes time—and emergence for that matter—to be real is excluded as impossible. Thus, Dennett like Laplace seems to use his metaphysical stance to assume what he wants to conclude. He assumes that reality across eternity could be given fully transparent and coherent. Hence, he claims, there were never any other possibilities.[11] In "the fullness of time" he writes above, only the actual history will exist. That might be so, but isn't it at least possible that even God only sees a partial truth? Could one imagine that some possibilities—or precursors—"existed" (in a non-reified sense), but never had any consequence[12] and thus are inscrutable, hidden even from the "fullness of time" perspective? This is a hard idea to entertain because we can never describe any such possibility without reifying and recognizing it, and thus installing it forever in the causal web. We might not therefore be able to conceive of this metaphysical idea from the metaphysical stance, but the classic formulation of the paradoxical stance has problems of its own and perhaps shouldn't be our measure of the possible.

D. Discrete and compositional determinism

Dennett does not just want to argue that only the actual course of history exists, he also wants to show that determinism is compatible with his notion of free will as evolving. Over the course of his books he presents different formulations of what

11. Dennett's position sounds and smells like a version of actualism, i.e. the position that only what is actual is possible. However, he vehemently argues that it is not—because he allows for lived and situated possibilities also interpreted in terms of capacities of objects. See (Dennett 1984, 144).

12. I shall not here be able to develop an emergentist or interactivist view of causality, but note that already Hume plays around with the idea that causing means producing, and that a cause can only be reified or understood through it effects, it is the "by which" (Hume 1750).

determinism is and I shall look at a few of them. In *Freedom Evolves* he starts out his discussion with Inwagen's definition that the claim of determinism is that "there is at any instant exactly one physically possible future" (Inwagen 1983, 3). He writes that this "not a particularly difficult idea." I beg to differ. As discussed in relation to Laplace, it is not clear how we are to actually think of "instants" that somehow contains something that fixes "physical futures"? Again a crux of my worry has to do with the timelessness of the material universe assumed under classical determinism, and the notion of a universe "time slices," "instants," "time t" etc.[13] We saw that Dennett himself talks about the notion of "time t" as a "mythical ideal" when it comes to the empirical study of the mind, however he seems perfectly content with it when defining determinism. The idea of an all-encompassing and extensionless time-slice of the arrested universe is in many descriptions of determinism operationalized as a "state description." Dennett writes: "A universe is *deterministic* if there are transition rules (the laws of physics) that *determine exactly* which state description follows any particular state description. If there is any slack or uncertainty, the universe is indeterministic" (Dennett 2003, 28–9). But interestingly Dennett quickly gives up on this definition as a useful tool for his argument. He writes: "There are too many fudge factors in this simple vision as it stands: How exact must a state description be? Must we plot every sub-atomic particle, and just which properties of the particle need to be included in the description?" (Dennett 2003, 29). I agree! The classical definition might not be feasible due to the challenge of pinning down the ultimate reality—or its "rules." Dennett of course sees this as merely a question of practical feasibility, and he then solves the issue by going a way from a material world model to a digital model without any "fudge." He writes:

> We can anchor these slippery factors arbitrarily by adopting another simplifying idea, W.v.O. Quine's proposal that we restrict our attention to simple imaginary universes, which he calls "Democritean" universes, in honor of Democritus, the most inventive of the ancient Greek atomist. A Democritean universe consists of some "atoms" moving about in "space." That is all. The atoms in a Democritean universe are not modern atoms full of quantum complexities but truly *a-tomic* (unsplittable, unsliceable) atoms, tiny uniform points of matter with

13. It is beyond the scope of this article to enter the deep debates within the philosophy of time, but I do want to motivate free will theorists to interrogate their typical treatment of the world as timeless. It is particularly ironic when the question at hand is how we as agent might effect *change* in the universe. For great discussion of the problems of the denial of temporality see Crome (2007) & Smolin (2013).

> no parts at all, rather like those postulated by Democritus. The space they inhabit must be made ultra simple, too, by *digitizing* it. (Dennett 2003, 29)

What has been described here? Dennett suggests that these "unfudged" imaginary all-or-nothing, uniform, ahistorical unsliceable atoms are a kind of "matter." But if so, in which sense? He also writes that they "move" in digitized space, but how so? Dennett has reduced the "universes" to sets of successive "states." In the quest to "unfudge" he seems to have defined away materiality, temporality and change. But do we not precisely need the fudge to portray a material and temporally changing world—i.e. is he loosing the material world he seeks to model in these first simplifying moves?[14] Dennett does not seem to think so. He uses this Quinean notion of Democritean universes to construct a "vast" but finite possibility space of "all possible worlds," namely all possible combinations of state descriptions in the digitized space given a predefined finite number of successive "instants." Pacing Borges, he calls this the Democritean Library. Each world represents one view from eternity, just simpler—and finite—as time and space are here explicitly discrete, and matter digitized. Given this finite Library of finite logically or "physically" possible and fully transparent universes, Dennett can now clarify Inwagen's definition. Determinism holds in the subset of logically possible worlds in which:

> "*There is at any instant exactly one physically possible future*. To say that determinism is true is to say that our actual world is in a subset of worlds that have the following interesting property: There are no two worlds that start out exactly the same (if they start the same, they stay the same forever – they are not different worlds at all), and if any two worlds share *any* state description exactly, they share all subsequent state descriptions." (Dennett 2003, 67)

Again, as we have already seen, the eternalist view does not pertain uniquely to deterministic universes but relies simply on the fact that everything is fully transparent and scrutable in each possible world. The news is that a "set" based definition of determinism is now not based on fixed transition rules, but rather on the notion that we browse the Democritean library, and compare and contrast "worlds." The very idea of a "world"

14. This is by no means a new insight but a paradox, which has been known at least since Parmenides and Zeno, and Democritus atomism was a compositional attempt to reconcile timeless being with our experience of time and change. See Crome (2007) for a fascinating analysis of the classic paradoxes and the conclusion that making sense of time and change might involve going beyond discrete math.

depends on the discerning intellect in the metaphysical stance, but now we are also asked to make "sets" of "possible worlds." This requires a meta-metaphysical stance, as I now need to step back from the individual world and imagine the whole vast Library of possible worlds to find out, which worlds have respectively overlapping and differing states. If two possible worlds after one overlap stay the same through eternity, then they are determinist worlds. With this definition the determinism—or indeterminism—of our actual world can be settled on the basis of sets of entirely discrete and unchanging digitized sequences, without any of "fudge" of describing or applying laws to create movement and change. In other words, the meta-metaphysical stance allows us to categorize across worlds and thus define determinism without ever introducing matter in lawful motion. With this toolbox Dennett's challenge is just to show that life can evolve in such "worlds", and this is where Conway's game of life comes into the picture.

E. Conway's game of life

One of Dennett's favorite models to show how complexity and avoidance can come from determinist underpinnings is Princeton Mathematician John Conway's digital "game of life."[15] The game consists of 1) a very simple finite two-dimensional world—a grid, with cells capable of being ON or OFF and discrete temporal instances. 2) A set of initial conditions where some and not all cells are ON - to get the "physics" going and 3) a set of very simple timeless and universal rules, which Dennett calls the "life Physics":

> Life physics: For each cell in the grid, count how many of its eight neighbors are on at the present instant. If the answer is exactly two, the cell stays in its present stage (ON or OFF) in the next instant. If the answer is exactly three, the cell is ON in the next instant whatever its current state. Under all other conditions the cell is OFF. (Dennett 2003, 36)

How does the transformation take place? How are the rules followed? From the perspective of each of the finite number of cells—we get an "input" of 1) number of ON cell neighbors and 2) ON or OFF status - which is categorically adjusted according to the rules of the "life physics." Note again that appearance and reality always neatly coincides. Dennett actually acknowledges this and writes "One of the delights of the Life world is that nothing is hidden in it; there is no backstage." It is a neat world of zero ambiguity,

15. See Gardner's (1970) for the original description in *Scientific American*.

where nothing exists, which is not recognized and responsive to the "life physics" at each instant.

Dennett, as mentioned, uses this simplified universe as an "intuition pump" for us to understand that life-like complexity can emerge from utterly deterministic rules. He points out that from an "intentional stance," we can recognize rather interesting patterns when we look at the unfolding changes of the "life world" from various distances and tempos. And he sees these higher-level patterns as in a sense real. Here is a quote from his discussion of pattern recognition in *the Intentional Stance*:

> Is the pattern that enables you to make the prediction "real"? So long as it lasts it is…The pattern may owe its existence to the intentions (clear-sighted or confused) of the machines designer, but its reality in any interesting sense – its longevity and robustness – is strictly independent of the historical facts about its origin. (Dennett 1987, 39)

The pattern is real and its reality is in its longevity and robustness. But what would that mean in the "Life world" as these patterns are indiscernable to the "life physics"? Is the reality in their availability to be recognized by *us* as external viewers? He writes:

> Whether one can see the pattern is another matter…I claim that the intentional stance provides a vantage point for discerning similarly useful patterns. These patterns are objective – they are *there* to be detected – but from our point of view they are not *out there* entirely independently of us, since they are patterns composed partially of our own "subjective" reactions to what is out there; they are the patterns made to order for our narcissistic concerns (Akins 1986). (Dennett 1987, 39)

Thus, pattern recognition is relational as it is "composed partially of our own "subjective" reactions."[16] The "interesting" patterns of the game of Life also clearly depend on us as viewers, the distance and speed by which we watch, as well as all the other recognition repertoires we bring with us (think Rorschach test). But is reification also relational at the "physical level"? It seems that the ON and OFF status of the cells must depend on their "recognition" as such, not by us but by the "life physics." In this

16. His description of the patterns and their objectivity and yet relational dependence can be seen as somewhat analogous to Gibson's view of affordances as objective but yet relative to the perceiver in their reification. See e.g. Gibson (1977).

connection it is curious that another of Dennett's favorite toy models is the "two-bitser"; a soda vending machine that recognizes and responds to certain coins and not others (Dennett 1987). Here he seems to propose, along the lines of Quinean inscrutability,[17] that if a two kinds of coins ("fake" and "real" quarters) cannot be discriminated, then there appears to be no further fact of the matter of which kind they are. The ultimate test is the actual interaction between coin and machine: if the soda comes out then it is a quarter.

In the two-bitser-case the interactions depend on a wealth of material details and complexities of the coins and the machine—which go beyond the recognition itself. I.e. no coins are actually identical material tokens and yet treated as such from the limited (and limiting) "perspective" of the machine. The material opacity has been eradicated in the game of life, where there is "no back stage" for either grid, cells or physics that goes beyond the recognition itself. The "Life worlds" are fully transparent two-dimensional worlds. Appearance equals reality and *every existent difference* is recognized and responded to with transition rules by the "physics," the omniscient Gods-eye-grid.

Dennett introduces the 'game of life' to show that complexity can emerge from strict physical determinism. But as even the "physical level" is digital it is not much like a material world in any traditional sense. There are not any persisting atoms nor any other "matter". Rather the cells go through instantaneous birth and annihilation without any component precursors or remains. Ex nihilo—in nihilo—like Leibnizian monads.[18] No continuous becoming and only something like instantaneous symbolic being. This digitalized all-or-nothing idea is necessary in Dennett's (and Leibniz') model to "unfudge" and stop the regress of further compositionality.

The question that I would like to raise is whether it is plausible that the ultimate level of universe would have any of these "life world" features? Note all the "unmoved mover" and "unregulated regulator" aspects of the Game of life. Where do the eternal

17. See e.g. Quine (1960).

18. Interestingly Leibniz monads are in many ways like digital pixels. See his monadology 1–6, in particular: "3. Now, where there are no constituent parts there is possible neither extension, nor form, nor divisibility. These monads are the true atoms of nature, and, in a word, the elements of things. 4. Their dissolution, therefore, is not to be feared and there is no way conceivable by which a simple substance can perish through natural means. 5. For the same reason there is no way conceivable by which a simple substance might, through natural means, come into existence, since it can not be formed by composition. 6. We may say then, *that the existence of monads can begin or end only all at once, that is to say, the monad can begin only through creation and end only through annihilation.* Compounds, however, begin or end by parts" (Leibniz, [italics mine]).

laws of physics come from? Where does the space grid come from, where do the pixels' coloring come from or disappear too? How are "instants" aligned, i.e. how are we to understand simultaneity or the idea of the "state of the universe at time t"? Dennett is a devout naturalist but yet these assumptions instill a finite border beyond which the "why" question does not apply. Dennett could here remind us that this model is purposively "unfudged." But as we have seen in our earlier analyses, it is not just Conway's model that introduces "unexplained explainers"—all assumptions of this "no back stage" collapse of appearance-reality collapse do. In other words, all Dennett's proposed metaphysical images do.

Beyond the question of the plausibility of "the buck stops here" assumptions of the framework, there is Dennett's question of whether the game of life could succeed in simulating evolution and freedom. In other words, returning to the issue of whether an unfudged model can show the complexity of life and pass the "Turing test" of life so to speak. Remember that in the two-dimensional discrete and determinist "Life worlds" all transitions are based on the "physical level" where appearance and reality never come apart. *We* can see the higher-level patterns, but the "life physics" cannot, and being external to the "Life world" our recognition is without consequence. In our lived world of epistemic agents and vending machines Dennett accepts the causal power of discrimination based on epistemic limitations. We are the loop. Dennett also insightfully proposes that whereas some interactions amplify variation, digitizing can be seen as a way of absorbing micro-variation. He writes:

> Surely, the result of a coin flip is the *deterministic* outcome of the total sum of forces acting one the coin…but this total sum has no predictable patterns in it. That is the point of a randomizing device like a coin flip, to make the result uncontrollable by making it sensitive to so many variables that no feasible, finite list of conditions can be singled out as the cause…*It accomplishes just the opposite of digitizing in computers: Instead of absorbing all the micro-variation in the universe, it amplifies it*… (Dennett 2003, 85 [italics mine])

This is an important point in regards to our metaphysical appearance and reality questions. Along these lines Conway's Game of life, and other digitized representations, like the two-bitser can be seen as owing their discrete all-or-none outcomes to condensing variation absorbing processes. In the case of Dennett's use of the Game of life as a metaphysical illustration, we are asked to disregard the physical computer, pretend there is no backstage, and limit our view to the "front stage" of the life world. We do

this by taking the deterministic laws, digitized "matter" and finite and discrete grids as "uncaused causes." But given Dennett's own analyses of relational categorization and pattern recognition would one not expect these digitized "Ex nihilo—in nihilo" creations and annihilations to precisely depend on a perspectival view? In other words, from the perspective of the two-bitser the quarter is either present or not. From the "grid-eye-perspective" of the "Life world" a pixel is either on or off. But like Dennett writes here such "unfudging" is generally the product of interactions absorbing variations. From a different perspective one might be able to discriminate two coins, which the two-bitser's recognition systems cannot, and thereby amplify consequences—e.g. by arresting people and so on. The point I am making is that each of Dennett's models gain their discreteness via the introduction of an ultimate "uncategorized categorizer," be it the God's eye point of view or the "grids" of ideal physics, or a giant immaterial two-bitser, in brief, the supreme perspective, beyond which no other discriminations or reifications can be made. Further, in determinist models like the game of life, causality is then restricted to this level of ultimate reification. Remember we can perceive, enjoy and name the patterns, but only influence them as "Deus ex Machina" by starting a new world. Thus, by way of an algorithmic Life physics-based occationalism the world is recreated at each instant given the prior reification.

I find it surprising that Dennett, on the one hand, notes that unfudging categorization is the *product*, the *output* of a variation absorbing process, and yet on the other assume an ultimate miraculously unfudged reality—be it determinist or indeterminist—in which "back stages" are per definition excluded as impossible. In other words, the question is why he does not pursue the possibility that a supreme causally-efficacious "uncategorized categorizer" might be impossible, not simply in practice due to our limited perspective, but due to the fact that reification needs a limiting perspective. It might be that Dennett is so focused on showing that compatibilism is *possible,* that life and freedom could conceivably have evolved from a determinist universe without "back stage," that he fails to consider whether this is a *plausible* story? At least this is what we see in his discussion of the "Game of life." He argues, based on Conway's proof, that one could embed a Universal Turing Machine in the life world. And therefore that a sort of self-replicating life *could* have evolved in a gigantic pixel space. He does worry about how to simulate variability, noise and mutation—i.e. re-create the fudge—in such a space: "Can a two dimensional world be noisy enough to support open-ended evolution, while still quiet enough to permit the designer parts to do their work unassailed? Nobody knows" (Dennett 2003, 50). He is right, we don't know. But we also don't know if God created us—and the fossil record. Some things might be possible but too unlikely to

spend time on, especially in the face of an alternative, that is. As it stands, I take it as a live option that there are no unmoved movers. We shall now look a bit at what it might mean if that were the case and we perhaps had interaction, hiding and revealing, and absorption and amplification all the way down.

4. Alternative views from within—emergence & interaction.

As mentioned Dennett suggests that his eternalist argument against possibilities works—ironically, for all possible worlds—and thus, whether the world is deterministic or indeterminist (Dennett 2003). In this way, he follows the free will debate consensus and takes classic determinism and indeterminism as the only metaphysical frameworks in town. Like Hume famously wrote: "As objects must either be conjoined or not, and as the mind must either be determined or not to pass from one object to another, it is impossible to admit of any medium betwixt chance and an absolute necessity" (Hume 1750). But is this true? Are necessity and chance, determinism and indeterminism our only metaphysical options? If we let go of the metaphysical stance assumption that the actual world has an ultimately reified causal level, then it seems that the possibility space would be open for other metaphysical options. One being that causality somehow takes place from within and exploits the gaps between appearance and reality. I shall in this article not attempt to develop such an alternative framework, but simply advocate for its possibility and briefly alert to the fact that I am not the first one objecting to the determinism-indeterminism ultimatum. Within the traditions of process philosophy, pragmatism, emergence and interactivist approaches many have pointed to metaphysical conceptions that does not fit the traditional eternalist metaphysical picture.[19] In a recent article Mark Bickhard defends the empirical plausibility of a process based interactivist and emergentist view. He writes:

> Process, in fact, is now the dominant language of science. Every science has progressed beyond an initial conception of its phenomena in substance terms to understanding that they are in fact process phenomena. Fire is no longer modeled in terms of the substance phlogiston, but instead in terms of the process of combustion; heat no longer in terms of caloric, but in terms of random kinetic processes;

19. See Vintiadis (2013) and Seibt (2012) for helpful recent overviews of respectively the emergentist tradition and process philosophy, but see also classic philosophy texts like e.g. Bergson (1896) and Whitehead (1929/1978), and of ecologists and theoretical biologists like e.g. Bateson (1979) and Rosen (1991).

> life no longer in terms of vital fluids, but in terms of special kinds of far from thermodynamic equilibrium processes. And so on. Every science, that is, with the exception of the sciences and philosophies of mind and persons. Here substance and structural views are still dominant. (Bickhard 2009, 553)

Thus, Bickhard points out how within nearly all sciences process-based and relational phenomena are taking center stage.[20] Interestingly, for our purposes here he is well aware that many metaphysicians have a hard time letting go of atomistic, determinist and eternalist assumptions and getting their mind around the notion of a process metaphysics.

> The shift to a process metaphysics, however, induces major changes in our overall framework of assumptions: First, change becomes the explanatory default, and it is stability that requires explanation. Similarly, processes, unlike atoms or the "stuff" of substances, do not have inherent boundaries, and boundaries too, therefore, must be explained, not assumed. Second, processes have their causal powers in virtue of their organization. Organization cannot be delegitimated as a possible locus of causal power without eliminating all causality from the universe. But, if organization is a potential locus of causal power, then so is higher level organization. In particular, there is no metaphysical block to the possibility of emergent causal power in new organization. And third, if emergence is a metaphysical possibility, then the door is open to the possibility that normativity and mind are emergent. (Bickhard 2009, 553–4)

In other words, to use another of Dennett's favorite expressions—we need to perform a "strange inversion of reason" to understand process metaphysics (Dennett 2003, 47). Of course Dennett takes the Darwinian inversion of reason to be a bottom-up compositional story, as opposed to a top-down story of intelligent design. What I, via Bickhard, propose is instead an inversion of reason away from reductionist foundationalism and "unexplained explainers" (be they large or small, physical or immaterial) to a view from within—all the way out so to speak.

20. See e.g. Witzany (2014) for recent insights in the "pragmatic turn in biology."

Such accounts can and have been developed in different ways, typically given differing critical starting points. Physicist Lee Smolin and philosopher Roberto Unger have in recent works (Smolin 2013, Unger & Smolin 2014) challenged assumptions about a-historical pre-given deterministic laws existing externally to the fabric of actual processes. If rather laws are inherent to and evolving with the concrete physical processes and conditions, then we get a radically different "view from eternity" than the one Laplace and Dennett avail themselves off. Smolin (2013) also points to several much earlier proponents of the 'laws evolve' view. Dirac writes in 1939: "At the beginning of time the laws of Nature were probably very different from what they are now. Thus, we should consider the laws of Nature as continually changing with the epoch, instead of as holding uniformly throughout space-time." Given Dennett's view of freedom as evolving, why would he not also entertain the idea that laws and "transition rules" might evolve as well? Why the externalist view of a world that comes pre-categorized? This is precisely the aspect of universal laws that Charles Sanders Pierce found the most problematic all the way back in 1891:

> To suppose universal laws of nature capable of being apprehended by the mind and yet having no reason for their special forms, but standing inexplicable and irrational, is hardly a justifiable position. Uniformities are precisely the sort of facts that need to be accounted for…Now the only way to account for the laws of nature and for uniformity in general is to suppose them results of evolution. (Pierce 1891)

This point could not be more pertinent to the issue at hand, and the metaphysical assumption of "unexplained explainers" that each of the formulations of determinism that we have met contains. Like Bickhard wrote above, given a process view it is stability and borders, not change, that need to be explained, Pierce similarly suggests that it is uniformities and laws that call for an explanation. Further, Nancy Cartwright has from a internal stand point of how actual science unfolds, advocated for an alternative metaphysical view, which she calls "the dappled world" (Cartwright 1999). The idea is here not only that our current laws of nature "lie," but that the world might simply be such that not even from the metaphysical stance are there universal and non-local regularities to be found. She has also worked on an Aristotelian inspired notion of "capacities" that could be congenial to an emergentist or "internalist" account of causation and thus an evolutionary view of action (Cartwright 1994).[21] Thus, even though as Bickhard writes

21. Cartwright is one the originators of the so-called Stanford school within the philosophy of science, which in

above the field of philosophy of mind seems to have stubbornly insisted on outdated and implausible metaphysical assumptions, alternatives do seem to exist—even if they are not neatly reified.

5. Conclusion: Evolving Dennett's story beyond determinism

Given Dennett's own analyses of the causal efficacy of epistemic limitations, hidden complexities and two-bitser inscrutability, his philosophy seems to show us beyond the external metaphysical stance. If there is no layer of reality without a back stage, where appearance always equal reality, then we might suggest that all causal interactions "amplify" certain variations and "absorb" others. In other words, Dennett's account is in many ways congenial to an internalist—or perspectival, interactivist and emergentist—inversion, were *all* causal effects might rely on some recognition and response. If the world—like the front stage of Conway's game of life—were such that all reality appears at each instant, i.e. reality and appearance never come apart, then this "internalist" idea of causality would make no difference. But imagine a world—perhaps like ours—with mind-boggling multiplicities of spatio-temporal timescales and ways of hiding and revealing. What if no fundamental level is to be found—but rather each level, each interaction has its own "dark matter" and even the "Higgs field" is not an "uncaused cause." Maybe in such a world there are no neat "states" and "instants," where everything can reveal all its aspects or properties. In such a world there would always be some gaps between appearance and reality, between what is of consequence in current interactions and what "it there." Such a temporal world of situated causalities I think would be more conducive to evolution, emergence, change and the odd combination of noise and quiet Dennett is looking to program into Conway's game of life.

If Dennett's compatibilism and his evolutionary account of free will are revisited with such considerations in mind then it looks as if parts of the possibility space have been ignored. Maybe a biological and historical view like Dennett's allows us to invent a new free will position that does not rely on traditional libertarian routes of claiming extra-physical determining forces (dualist) or non-determined (random) effects. The question is whether a dynamic pluralistic world always getting its causality from actual interactions, with limited relational recognition, can provide the grounds we need to re-interpret how real-time action choice and determination might be possible at the human level. More specifically, whether the causal consequences of teleological perception,

many ways is united not simply by geography but by its challenge to eternalist, essentialist and reductionist views. See e.g. Dupré (1995).

memory, anticipation and interaction feedback, can give rise to forms of emergence that we might think of as having downward or—as I prefer—outward efficacy.[22]

Some might want to argue that this proposal is not actually that different from Dennett's, as he might be read not necessarily as a classic determinist but as a metaphysical agnostic. This might be so, but then this article is making a significant contribution to the clarification of Dennett's position and its commitments and explanatory merits. His texts are, as we have seen, ripe with reductionist and eternalist models, which indicates that he at least in practice is ignoring other metaphysical possibilities such as emergentist, interactivist and process positions.

I propose that the possibility of such positions matter—also to our conception of our own actions. Instead of saying, as Dennett does, that we have perceived opportunities because we are epistemologically limited, we can now say that our epistemological limitations are *as casually efficacious as anything else can claim to be*. The point is that in a world that cannot be reduced to a level of ultimate and inherently transparent compositional "parts," *every* interaction might be only partially revealing. The world of epistemology and metaphysics, of appearance and reality, phenomena and noumena, now interacts as no processes of reification are externalized from the causal web. This might allow for a changed perspective on how our action choices matter. The question is if our evolved abilities to predict, and flexibly integrate horizontal cultural influences, and inhibit and hide impulses, are the sources of emergent effects in the present, i.e. shape the causal fabric of the world in ways that are as determining as they are determined. The proposal is thus not that our actions are "indetermined" in any classic sense but rather that they both—to use Dennett's terminology—*amplify and absorb variability*. We can actively and purposively *hide* and *reveal* causally efficacious aspects of "ourselves" and the world.[23] Dennett is right to highlight our complex and bastardized genealogies, as this heterogeneity and multiplicity of the self holds "it's" generative power. It is not just that we "are the loop" as Dennett says, but further that we constantly dynamically change and choose what pertains to *our* causally efficacious loops. Dennett aims to show that freedom can evolve from entirely unfree atomistic parts, and accordingly likes to describe

22. An important aspect to explore of such a hypothesis would be whether in a world without a bottom layer, one might expect both order and disorder—regularity and variability—to be constantly be produced as irreducible actual concrete interactions would determine the outcomes in real time. This does not seem to be the case in Conway's game of life—here the finite determinist worlds move towards more order.

23. Note here the similarity to Merleau-Ponty's (1964) the insistence on the role of invisibility in visibility as well as his critique (1965) of what he sees as Bergson's (1896) purely positive process philosophy.

us in mechanistic terms as composed of "micro-robots" and "micro-factories." However, we might—perhaps like the bacterium—in part owe our freedom to not being the sum of a fixed set of finite parts, but rather constantly self-creating open systems. Thus, autonomy and emergent novelty might be possible.

I want to reiterate that I do not claim to have a theory of free will. My aim in this present an article is merely to reveal theoretical possibilities that have been hidden by the insistence on fully transparent discrete and eternalist metaphysical perspectives. If there is no floor to the universe, then there might always be a gap between appearance and reality. This would undermine all classic formulations of determinism it seems. However, I tried to show that the potential absence of a metaphysical "back stage," suggest that our action choices might indeed happen in real time and determine what is to come, possibly by way of new emergent effects. Thus, the elbow room of human agency comes not only from the fact that our procedural selves are "smeared over time and space" but also from the constant games of peek-a-boo where our worlds and we—in all our bastardized multiplicities—are never fully hidden nor fully transparent or expressed.[24] These ideas might allow for the development of an ignored kind of free will worth wanting, and I shall be curious to see if some indeterminists would agree. After all, as Dennett is, I am puzzled by the desire to have one's actions to be without any precursors. And, I wonder whether many do not simply desire—as James (1897) famously expressed it—"that the issue is decided nowhere else than here and now. That is what gives the palpitating reality to our moral life and makes it tingle as Mr Mallock says, with so strange and elaborate excitement."

24. Bergson - and other who have made similar claims - might thus exactly be wrong to propose that the free action was an expression of the entire 'self'. As he his self later pointed out; all actions simplify. (Bergson 1889)

References

Akins, Kathleen. 1996. "Of sensory systems and the "aboutness" of mental states." *The Journal of Philosophy* 93 (7): 337–372.

Bateson, Gregory. 1979. *Mind and Nature—a necessary unit*. New York: Bantham New Age Books.

Bergson, Henri. 1889. *Essai sur les données immédiates de la conscience*. Paris: F. Alcan.

Bergson, Henri. 1896. *Matière et mémoire*. Paris: F. Alcan

Bickhard, Mark. 2009. "The Interactivist model." *Synthese* 166 (3): 547–591.

Brincker, Maria. 2014. "Navigating beyond "here & now" affordances—on sensorimotor maturation and "false belief" performance." *Frontiers in Psychology* 5 (1433). doi: 10.3389/fpsyg.2014.01433.

Cartwright, Nancy. 1994. *Nature's Capacities and their Measurement*. Oxford: Oxford University Press.

Cartwright, Nancy. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Crome, Victor J. 2007. *Zeno's Paradoxes and the passage of time*. Doctoral dissertation. New York: City University of New York.

Dennett, Daniel C. 1984. *Elbow room: The varieties of free will worth wanting*. MIT Press.

Dennett, Daniel C. 1987. *The intentional stance*. Cambridge: MIT press.

Dennett, Daniel C. 2003. *Freedom evolves*. New York: Viking Penguin.

Dirac, Paul. 1939. "The relation between mathematics and physics." In *Proc. Roy. Soc. Edinburgh* 59 (II): 122.

Dupré, John. 1995. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge: Harvard University Press.

Gallagher, Shaun. 2006. "Where's the action? Epiphenomenalism and the problem of free will." In *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition,* edited by W. Banks, S. Pockett, and S. Gallagher, 109–124. Cambridge: MIT Press.

Gardner, Martin. 1970. "Mathematical Games—The fantastic combinations of John Conway's new solitaire game "life."" *Scientific American* 223 (October 1970): 120–123.

Grossman, Lisa, and Richard Webb. 2013. "Martian chemistry was friendlier to life." *New Scientist* 219 (2933): 14.

Hornsby, Jennifer. 1997. *Simple Mindedness—In defense of naïve naturalism in the philosophy of mind*. Cambridge: Harvard University Press.

Hume, David. 1750. *An enquiry concerning human understanding*.

Inwagen, Peter van. 1983. *An essay on free will*. New York: Oxford University Press.

Laplace, Pierre-Simon. 1951. *A Philosophical Essay on Probabilities*. Translated by F.W. Truscott and F.L. Emory. New York: Dover Publications.

Leibniz, G. W. (1893) 1989. "The monadology." In *Philosophical Papers and Letters*, edited by Leroy Loemker, 643–653. New York: Springer.

Merleau-Ponty, Maurice. 1964. *Le Visible et l'Invisible*. Paris: Gallimard.

Merleau-Ponty, Maurice. 1965. *Éloge de la philosophie*. Paris: Gallimard.

Millikan, Ruth Garrett. 1989. "In defense of proper functions." *Philosophy of Sscience* 56 (2): 288–302.

Peirce, Charles Sanders. 1891. "The architecture of theories." *The Monist* 1 (2): 161–176.

Putnam, Hilary. 1981. *Reason, truth and history.* Vol. 3. Cambridge: Cambridge University Press.

Quine, W.V.O. 1960. *Word and Object*. Cambridge: MIT Press.

Rosen, Robert. 1991. *Life Itself—a comprehensive inquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.

Seibt, Johanna. 2012. "Process Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Smolin, Lee. 2013. *Time reborn: From the crisis in physics to the future of the universe*. Boston: Houghton Mifflin Harcourt.

Unger, Roberto Mangabeira, and Lee Smolin. 2014. *The singularity of the universe and the reality of time.* Cambridge: Cambridge University Press.

Vintiadis, Elly. 2013. Emergence. *Internet Encyclopedia of Philosophy*.

Whitehead, Alfred North. (1929) 1978. *Process and reality*. Corrected eds. David Ray Griffin and Donald Sherburne. New York: Free Press.

Witzany, Guenther. 2014. "Pragmatic turn in biology: From biological molecules to genetic content operators." *World Journal of Biological Chemistry* 5 (3): 279–285.

Journal of Cognition and Neuroethics

Free Will and Autonomous Medical Decision-Making

Matthew A. Butkus
McNeese State University

Biography

Dr. Matthew A. Butkus earned his undergraduate degrees from Georgetown University (German and Philosophy) and the University of Pittsburgh (Psychology). He earned his graduate degrees from Duquesne University (MA in Philosophy and PhD in Health Care Ethics). He is currently an Associate Professor in the Department of Social Sciences at McNeese State University. His prior academic and clinical appointments include Chatham University, Mercy Hospital–North Shore Campus, the CRISMA Research Group in the Department of Critical Care Medicine at the University of Pittsburgh Medical Center (Clinical Research, Investigation, and Systems Modeling of Acute Illness), and St. Francis Health System.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Butkus, Matthew A. 2015. "Free Will and Autonomous Medical Decision-Making." *Journal of Cognition and Neuroethics* 3 (1): 75–119.

Free Will and Autonomous Medical Decision-Making

Matthew A. Butkus

Abstract

Modern medical ethics makes a series of assumptions about how patients and their care providers make decisions about forgoing treatment. These assumptions are based on a model of thought and cognition that does not reflect actual cognition—it has substituted an ideal moral agent for a practical one. Instead of a purely rational moral agent, current psychology and neuroscience have shown that decision-making reflects a number of different factors that must be considered when conceptualizing autonomy. Multiple classical and contemporary discussions of autonomy and decision-making are considered and synthesized into a model of cognitive autonomy. Four categories of autonomy criteria are proposed to reflect current research in cognitive psychology and common clinical issues.

Keywords

Cognitive psychology, neuroscience, medical ethics, autonomy, heuristics, backstage cognition, decision-making

Since its inception, medical ethics has concerned itself with balancing several key concepts—the patient's best interest, both psychosocial and medical; the patient's legal rights and autonomy; the authenticity of the patient's decision, i.e., narrative concerns that the patient's choice be reflective of her values, etc. As is the case with any pluralistic system, these concepts are complementary at times and conflicting at times. Significant efforts to determine just how to proceed in any given case result, both in academic circles, in which theories clash and value structures rise and fall, as well as in clinical cases, in which academic language gives way to clinical context and lives hang in the balance.

These concepts of autonomy and authenticity have dominated ethical thought for several decades, and have been given significant, if not complete, weight in many theories. Autonomy is seen by many as a deontological norm—an absolute right and duty in some models, a *prima facie* duty in others. Its value and moral weight are understood as being *a priori*—it is not contingently valuable or worthy simply as a means to some other end. The purpose of this paper is to explore this concept of autonomy, and to see how it is modified by knowledge from multiple fields.[1] Philosophy certainly offers

1. I first explored this critique of autonomy in light of depressive illnesses and the decision to forgo medical treatment in my doctoral dissertation (see Butkus, M. A. (2006). *Depression, Volition, and Death: The*

compelling accounts and definitions, but a fundamental question arises: what does the concept mean in light of what we have learned from fields like cognitive psychology and psychiatry? Philosophy and ethics have debated 'modifiers of the voluntary' for a long time, but these concepts of coercion generally are predicated on conscious awareness and experience.

A more complete model of cognition notes that significant thought processes occur at levels which we are only beginning to understand. These influences are non-conscious: they stem from a collection of processes outside of our conscious awareness. How, therefore, can we exercise control over or appreciate the influence of elements of which we aren't aware? Many models fiat the ability of the moral agent to choose amongst alternatives—these models seem to be less compelling in light of what we know and understand from other disciplines. In fact, the more we learn about the brain, the more homuncular they seem—it is almost as if they argue for a little man sitting in our brains, selectively choosing what will influence us to act. These models are untenable—any conception of autonomy must include an appreciation for cognitive elements outside cognition, which potentially bias us in ways that are inauthentic. In upholding choices that may be skewed or inauthentic, we undercut any meaningful sense of autonomy. A proper sense of autonomy, therefore, is much more deterministic and less 'rational' than modern models suggest. As such, greater care is necessary in assessing competence to forgo treatment—quite simply, current popular models allow for more bad decisions with fatal consequences, a reality antithetical with the stated and implied purposes of ethics in medicine. We destroy that which we would protect in a decision which may be the last choice the patient ever makes. If we genuinely care for our patients, we ought to help them reach meaningful choices, instead of fiating an empty and ill-defined autonomy.

Case Study

William R. is a 45 year-old man with end-stage renal disease. He is dialysis-dependent and requires treatment three times per week. In his last hospitalization, he explained to his treatment team that he no longer desired to receive dialysis, maintaining that he felt

---

Effects of Depressive Disorders on the Autonomous Choice to Forgo Medical Treatment. Pittsburgh, PA: Duquesne University. doi:10.13140/2.1.3236.9284). That research and analysis strongly informs this work. Recent arguments in medical ethics have also explored the impact of mental disorders and their implications for the free will debate (see Meynan, Gerben. 2010. "Free will and mental disorder: Exploring the relationship." *Theoretical Medicine and Bioethics* 31: 429–443.; Müller, Sabine, and Henrik Walter. 2010. "Reviewing Autonomy: Implications of the Neurosciences and the Free Will Debate for the Principle of Respect for the Patient's Autonomy." *Cambridge Quarterly of Healthcare Ethics* 19: 205–217).

it would be too burdensome for him to continue. An ethics consult was called, and the consult team met with the patient for over an hour, discussing his understanding of what forgoing treatment would mean and his reasons for electing not to continue dialysis. He described his personal history, in which friends of his became dialysis dependent and were unable to continue with their hobbies and personal interests. He disclosed that he found the treatments prevented him from eating out, socializing, and enjoying other activities. He understood that absent his treatments, his physical state would deteriorate, culminating in his death in a matter of weeks.

Throughout his interview, the consult team did not find any immediate reason why he could not exercise his personal autonomy and forgo treatment. The consult was written up elucidating the reasoning for supporting the patient's decision and the team's recommendations were followed. William died the following week.

Critiques of Autonomy and Classical Models

William's case is an example of a classic issue in medical ethics—the ability to forgo medical treatment is quite likely *the* most common and accessible example of medical ethics to the general public, with representations in popular television, movies, and other mass media. A patient's ability to act autonomously is rightly praised—individual liberty is highly valued in Western society and if our ability to act is going to be curtailed, we require a significant level of justification for doing so.

This is not to suggest that autonomy is not without its detractors—the question is raised as to whether we have overcorrected from the paternalism of the past, in which physicians would routinely substitute their own preferences for those of their patients. Autonomy has been criticized from feminist philosophy and sociological viewpoints,[2] for

2. Donchin, Anne. 2001. "Understanding autonomy relationally: Toward a reconfiguration of bioethical principles." *Journal of Medicine and Philosophy* 26 (4): 365–86; Homan, Richard W. 2003. "Autonomy reconfigured: incorporating the role of the unconscious." *Perspectives in Biology and Medicine* 46 (1): 96–108; Jennings, Bruce. 1998. "Autonomy and difference: The travels of liberalism in bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 258–69. Upper Saddle River: Prentice Hall; Lane, Robert E. 2000. "Moral blame and causal explanation." *Journal of Applied Philosophy* 17 (1): 45–58; Light, Donald W., and Glenn McGee. 1998. "On the embeddedness of bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 1–15. Upper Saddle River: Prentice Hall; Parks, Jennifer. 1998. "A contextualized approach to patient autonomy within the therapeutic relationship." *Journal of Medical Humanities* 19 (4): 299–311; Roessler, Beate. 2002. "Problems with autonomy." *Hypatia* 17 (4): 143–62; Tauber, Alfred I. 2003. "Sick autonomy." *Perspectives in Biology and Medicine* 46 (4): 484–95; Wolpe, Paul R. 1998. "The triumph of autonomy in American bioethics: a sociological view." In *Bioethics and Society: Constructing the Ethical*

instance, and arguments for softer forms of paternalism have been recently proposed (Conly 2013). Despite these critiques, autonomy remains highly valued in both law and philosophy, and we still maintain a high standard for valuing other principles above it. The issue at hand is whether this standard is artificially high—that is, is it higher than warranted given more naturalistic explorations of the phenomena of consciousness and decision-making. Does it make sense for us to reappraise both the value of autonomy as well as the criteria defining it?

Classically, autonomy has been linked with human reason—for theorists like Aristotle, Kant, and Descartes, reason and rationality are essential defining characteristics of humanity. This classic model of autonomy involves five assumptions about rationality, including literalness, logic-dependence, conscious experience, disembodied transcendence, and essential emotional disconnectedness. Contemporary research has challenged or debunked these assumptions, yielding a model of rationality that is dependent on metaphor, metonymy, inferential reasoning, and unconscious processing, and which is fundamentally connected to and influenced by emotion (Lakoff 1999). Further, this model is known to skew perceptions of new evidence, to have limits in scope, and to be contextualized (Evans & Hollon 1988; Miller & Moretti 1988). Ultimately, this empirical cognitive model defies rationalist claims. This does not make it easy, however, to abandon classic notions of radicalized autonomy—there is still a visceral appeal to the idea that I am in full control of my thought process and the actions that result from it. However, if we want to be honest and move towards a sense of autonomy that matches up with the available data, we must become much more aware of the role of the unconscious and backstage elements of our cognition. Continuing to insist that medical autonomy reflect classical and rationalist models of cognition is dangerous—it promotes an ideological model divorced from actual decision-making. Human thought is much more complex, reflecting deductive, inductive, and abductive reasoning influenced by unconscious and backstage elements *and* streamlined by a number of cognitive heuristics hardwired by evolution. The recognition of these unconscious backstage elements has required a reimagining of the concept of freedom and autonomy (Hájíček 2009; Levy 2003; Shepherd 2012).

Cognition is not a single-stage process—there are many levels of organization in the brain, and they interact with each other in many ways which are open to influence. Conscious thought—the result of myriad physical and social interactions, is also a construct; a concatenation of many different types of cognition, operating in conscious

Enterprise, edited by Raymond DeVries and Janardan Subedi, 38–59. Upper Saddle River: Prentice Hall).

and backstage capacities. Backstage cognition involves a variety of related concepts, e.g., reflexive thought patterns with affective and behavioral components, generation of novel meanings for situations and objects from the mental assembly of other situations and objects, and distinctions between algorithmic and heuristic thought. Our concern is not with the conscious elements of cognition, as conscious phenomena are predicated on deeper phenomena. We cannot have conscious experience without deeper structures, much as we cannot build a castle before constructing its foundation. All of the myriad sense data we take in initiate complex activation pathways, associating current stimuli with previous experiences, affective data, and other valence structures. These deeper cognitive phenomena are not simplistic processes—they are layered, quite complex, exceptionally fast, and quite independent of our volition (Ashcraft 1994). Their automaticity belies their complexity—just as complex physical responses can be initiated without volition, so too we should recognize that our cognitive processes can be induced to action. An environmental trigger can give rise to the activation of many complex systems—a particular memento can trigger complex memory and affective components with corresponding behavioral components (Smith 1997). For instance, I may pass a photograph of my grandfather, which triggers a series of memories (living with my grandparents, visits, holidays, advice given to me, etc.), eliciting specific affective responses (sorrow at his passing and resolution to fulfill promises made to him), and culminating in actions and behavioral changes. None of these responses were necessarily *chosen* by me—they are all direct results of the environmental stimulus; further, this same stimulus can affect me well after I actually encounter it—my *memory* of the stimulus can provoke identical psychological and behavioral responses.

What is more, these backstage processes are also able to introduce errors into cognition—the way we perceive the world is dependent upon a variety of factors, some within our control, some well outside control. A requisite part of accurate cognition is appreciating and understanding when we are making choices based upon the indeterministic elements within our control and the deterministic elements lying outside our volition or awareness.

Automaticity is a significant element of cognition—a variety of processes simply occur without volitional cueing.[3] Bargh understands automatic cognitive processes

3. The simplest means of demonstrating this is by asking the question "What is the first thing you think of when I say the words 'white bear'?" The normal reaction is to call to mind immediately an image of a polar bear—this was not a voluntary process, however, in that had the words pointed to some other cognitive target, you would be free to think of myriad other things instead of white bears.

to occur "reflexively whenever certain triggering conditions are in place; when those conditions are present, the process runs autonomously, independently of conscious guidance" (Bargh 1997). This can refer both to physical processes—such as navigating an automobile while thinking of something entirely different—as well as cognitive processes—such as references to white bears cueing the imagination of polar bears. Isen and Diamond clarify Bargh's model, noting that automatic processes are best understood as a 'parallel process'—they do not take up cognitive processing resources (attention or effort), so they can occur parallel to other cognitive processes which do require these resources (Isen 1989). Because it does not tax cognitive resources, automatic processing can be performed much more rapidly and earlier than other types of processing. This may explain our 'gut instincts' in certain situations—our full processing has not yet finished, leaving us with only a general impression of necessary action. Berkowitz notes that the deterministic model suggested by automaticity is frequently undervalued by many people—there is a frequent visceral objection to the idea that our cognitive processes are heavily influenced by environmental determinants (Berkowitz 1997). These can be manifested as objections to experimental results or methodologies or as appeals to the indeterministic claims of folk psychology. Berkowitz suggests that, if nothing else, "Persons interested in gaining a truly adequate understanding of the complexities of human conduct should at least adopt a healthy skepticism toward the assumption that conscious processes are necessarily involved in all human behavior" (Berkowitz 1997, 85). As much as the average moral agent would like to dismiss them, unconscious and preconscious processes can be powerful determinants, and not just modifiers, of the 'voluntary.'

Preconscious processes develop as the result of conditioning—we develop patterns of psychological responses to stimuli. As is claimed by behaviorist thought, we make associations between stimuli and psychological responses, facilitating future responses along those same psychobehavioral lines. It becomes easier for stimuli to elicit behavioral, emotional, and motivational responses in us, producing automatic cognitive processing. Initially these responses require work, but like other recurring responses, the amount of conscious effort they require consistently decreases to the point where they require no conscious processing at all (Bargh 1997). This has serious ramifications—it means that if we encounter a particular cognitive trigger, we can initiate goals, motivations, and resultant behaviors automatically. Absent volitional control, we may not necessarily be able to control the kinds of thoughts and actions that result. In a clinical setting, for instance, a particular diagnosis may be an emotional trigger for a variety of subsequent thought processes and associations. The mere word 'cancer' may elicit a slew of memories

and experiences involuntarily and instigate thought processes culminating in a comorbid depression, which may radically affect how our patient perceives his or her current health and prognosis. When asked about treatment preferences, and whether the patient desires a particular course of treatment, we may have unknowingly set into action an automatic process that results in an outcome our patient might not otherwise desire.

In contrast to this proposed highly deterministic model, Baumeister and Sommer suggest that consciousness introduces explicitly indeterministic elements (Baumeister & Sonner 1997). They argue that consciousness allows us to recognize when automatic processes are occurring, and to exercise control in the behavioral process. Introducing some indeterminacy into decisional models does not contradict underdetermined decisional models, and it allows for ownership of action with accompanying ethical valence (moral praiseworthiness or blameworthiness). It reinforces the necessity of exploring the decisions we make to ensure that they are, in fact, the result of conscious mediation, and not simply the result of underlying automatic processing. I wish to stress that there are *strongly* deterministic causal factors in cognition, and that we must be aware of the myriad influences upon our choices, especially in critical situations such as forgoing treatment.

Automaticity, therefore, can be a powerful motivator for action, resulting in affective changes, goal activation, and deterministic mediators of conscious processes. These resultant changes are necessarily interactive and modifying causal elements of further cognition. As a result, we see that cognition has strongly deterministic elements at all levels of pre- and post-conscious processing. These elements necessarily conflict with our folk model of cognition, in which our decision-making is essentially free.

As such, the model that emerges from this discussion is that of a consciously mediated but often deterministic, reflexive processing in response to both external and internal stimuli which can have long term effects on affect, perception, and cognition. In short, the choices that we make can be heavily influenced, *but not necessarily determined*, by factors outside of our control. Clinicians should be very aware of the role that context and psychological stimuli have upon the decision-making process. If a patient chooses to forgo medical treatment, we would be remiss if we were not to ensure that it is done for the right reasons, and not as an automatically processed reaction to the situation in which the patient finds him or herself.

The discussion of cognition must also contain a discussion of 'mental spaces' and 'backstage cognition'—a theory of cognitive processing positing the assemblage of novel ideas and constructs from earlier ideas and constructs, occurring outside of our conscious awareness (Fauconnier 1994). Fauconnier argues that language cues give rise to cognition

outside of our awareness, building complex cognitive structures that can exceed the extent of the information presented. He suggests that *any* form of thought or cognition produces such mental spaces, and stresses that these ought not to be considered simulations of reality of 'possible worlds'—consequently, we ought not to envision them as such or compare them to types of heuristics setting up simulations of possible outcomes. These elements, however, are not necessarily accessible to us consciously—we are engaging in a phenomenon called 'backstage cognition'.

The cognitive processes of which we *are* aware are surface phenomena, and merely a subset of all the phenomena occurring when we consider choices and options. Thought and judgment are much more complex processes than our everyday folk accounting would suggest, and any model of 'rational autonomy' must account for a profound empirical criticism—'rationality' isn't so rational after all. This is a very different model than what we encounter in classical models of moral agency, which posit a decision-maker as rationally mapping out the consequences of particular actions and assigning objective probabilities to each. Instead, cognition appears generally to be more *ad hoc*—judgments and meaning seem to be constructed by conceptual blending in mental spaces, rather than the results of conscious deliberation.

The material that is drawn into the blend does not have to be part of the current stimulus—it is entirely possible for one to draw upon old experiences and memories as inputs into a conceptual blend. This will be an important part of the cognitive autonomy model as well—experience and memory provide the information accessed most readily, in addition to emotional valences. We are not necessarily aware of all of the blends that our minds produce—as it is a backstage process, it is entirely possible for meanings and associations to be blended, but to be preconsciously rejected in favor of other interpretations (Fauconnier 2002). They may be rejected for a variety of preconscious reasons; while we do not presently have a full accounting of preconscious processes or reasoning (and, in light of our complexity, one might reasonably ask whether we will *ever* have such an account), we have several candidate theories in heuristics-and-biases, ecological rationality, bounded rationality, and 'fast and frugal' heuristics.

In essence, the way we think about many things is not necessarily based on the strongest information or the most accurate understanding of what information we *do* choose to focus on. Further, we are often called upon to evaluate novel situations, and in this context, we find that there are several typical constraints upon what we view as likely versus unlikely, based upon any germane or potentially relevant information we possess. We construct scenarios to evaluate how we can reach the targeted outcome; the more plausible the scenarios we discern, the more likely the target event. In principle, this

serves as a common standard to decide between distinct alternatives, appreciation of the consequences of these choices, and the process culminating in the choice that maximizes the return the agent receives as measured by the common standard.

In a clinical setting, this is a description of our idealized patient and our ideal of informed consent—authentic choices predicated on an understanding of the procedures and risks involved and knowledge of the reasonably predictable outcomes. There is a problem, however—this standard is impossible. We have innate limitations on how much information we can manage in constructing these scenarios; as a consequence, we tend only to alter simple elements or factors, which may not conform to reality or may be counterintuitive (Tversky & Kahneman 1982). Further, once we construct a particular scenario, we tend to find it difficult to imagine other possibilities—we become tied or 'anchored' to one given possible explanation or course of action, which limits our ability to generate further scenarios or to see other potential outcomes. Tversky and Kahneman further note that in judging probabilities and unknowns, our decisions are only adequate if the judgment is in accord with the entire collection of beliefs held by the thinking agent. This poses a problem in assessing rationality: there is no simple way to check whether any particular set of probability judgments are compatible with the individual's collective whole. Instead, the individual simply strives for conscious and unconscious compatibility with his knowledge, assessments of probability, and his own heuristics and biases. In other terms, the individual strives to make his decision as authentic as possible.

Further modifying our knowledge pool complicates our decisional framework—we respond differently when we begin to add information into our cognitive schema. Our mind can have difficulty filtering useful information from worthless information—studies demonstrate that "people respond differently when given no evidence and when given worthless evidence. When no specific evidence is given, prior probabilities are properly utilized; when worthless evidence is given, prior probabilities are ignored" (Tversky & Kahneman 1982, 5). When information is present, we assign it decisional weight and importance, but may potentially give it undue weight, leading us to become either overly reliant upon that particular piece of information (anchoring), or overly confident in our assessment of its worth, a failure rampant across lay and professional decision makers.

Human cognition does not follow an overtly rational process like pure information processing and utility maximization; our cognition is characterized by values, emotions, prior knowledge, raw intelligence, and many other factors that do not fit nicely into this idealized model. Accounts or theories of autonomy must reflect this messiness to be sound—if our philosophy is not influenced and tempered by what we learn from neuroscience and cognitive psychology, it is an exercise not in truth but in fiction (Lakoff

1999). Special interest in concepts like backstage cognition and heuristics and biases can be traced back decades (Ashcraft 1994; Gigerenzer 1996; Gigerenzer, Czerlinski, and Martignon 2002; Gilovich & Griffin 2002; Kahneman 2011; Tversky & Kahneman, 1982), but medical ethics has not broadly integrated these findings. Models of medical autonomy from that period evidence a classical understanding of rationality and reason, and three principle models serve as examples.

Homuncular Autonomy Models

Some of these models explicitly endorse the classical cognitive model, while others only make covert appeals by linking biases and thought distortions to psychopathologies or outside influences. They propose a model of cognition which seems to suggest a high (if not total) degree of control over what influences us in our thought processes, with our only weaknesses being disease, addiction, immaturity, or dementia. The evidence of the past few decades of research in cognitive psychology and neuroscience paint a very different picture.

Veatch

The first model of note from that era comes from Robert Veatch, in which he establishes a relationship between deontological and consequentialist methodologies and principles, producing a system advocating promise keeping, beneficence, and personal autonomy (Veatch 1981). Veatch is especially concerned with liberty rights—a category of claims that prevents others from infringing upon our ability to act. Contained within this category are the right to refuse treatment and the right to control one's body. Related to this is the ability to act on the information disclosed by physicians—Veatch defends a scenario in which giving a patient unwanted information constitutes as much of an ethical violation as failing to provide information. In essence, Veatch defends autonomy rights over the provision of information, allowing for a model in which the physician must respect the autonomous decision-making of a person even in the face of obvious ignorance of salient facts. Veatch explicitly makes patient autonomy a trump—we have a duty to respect it at all costs and even in circumstances when there is sufficient justification for questioning it (e.g., psychiatric hospitalization). This need to respect autonomy extends to patients in a variety of circumstances, including those who may be experiencing terminal illnesses, which potentially impacts or compromises their ability to make decisions. Veatch recognizes that one's autonomy and moral decision-making do not exist in a vacuum—his model recognizes that the patient's moral community includes

other relationships that must be factors into decision-making. This allows an outward growth of our understanding of personal autonomy, but not inward growth into our own thought processes.

Autonomy is a deontological norm in Veatch's model—it is seen as a prerequisite for evaluating a moral action (per his argument, we must address our prima facie duties before attending to their consequences). We cannot justify violating a nonconsequentialist principle, regardless of the good consequences our action may produce. This philosophy is understandable; it is entirely possible to imagine circumstances in which good consequences result from obviously immoral actions (e.g., peace produced by genocide). A moral system allowing for such an outcome is obviously suspect at best.

Despite this intuitive appeal, there are significant challenges to the autonomy concept as proposed by Veatch. Fundamentally, the picture of autonomy he proposes is built on an unrealistic cognitive model—he allows for illness to occasionally compromise a patient's competence (e.g., delirium), but he is more concerned with considering exceptions resulting from a patient's lack of information (in essence emphasizing the informed aspect of informed consent over the consent aspect). This is a clear deficit, as has been explored in the past few decades—we know much more about how the brain functions at a variety of levels of organization (from individual neurons to neural networks). We know that pathophysiology impacts our brains at the cellular and functional levels. We know that psychopharmacology and psychoneuroimmunology introduce additional factors to our unconscious thought processes. We have found any number of cognitive "rules of thumb" that creates shortcuts in decision-making that operate at levels we do not control. All of this creates a cognitive model far removed from what Veatch proposed. Obviously it is wrong to criticize a historical system based on recent findings, and much of the relevant work in cognitive psychology and neuroscience postdates Veatch's proposal. His argument, however, introduces a larger and recurring discussion in medical ethics contemporaneous to and following our insights into how we actually think.

Faden and Beauchamp

The second noncognitive model is that of Ruth Faden and Tom Beauchamp (1986), first published five years after Veatch. They also stress the essential (if not primary) importance of autonomy in medical ethics, defining it in terms of individual rights, and the obligations we have not to infringe on the ability of others to act. They include a

variety of concepts under the umbrella of autonomy, including privacy, voluntary decision-making, and accepting the consequences of one's decisions. This position strongly reflects the root of autonomy—the principles of self-governance and self-direction. Faden and Beauchamp focus significantly on outside factors that can impact decision-making, especially the clinical staff (e.g., withholding information relevant to the treatment decision, not recognizing the patient's refusal of treatment, etc.). Just like Veatch, they place autonomy into a pluralistic system in which multiple values are weighed in ethical decision-making. Unlike Veatch, however, they do not give autonomy trump power—they envision circumstances in which beneficence and justice require us not to respect the patient's autonomy.

Their picture of autonomous agency does not posit a variety of strict criteria. They focus on a model of autonomy that meets our everyday understanding and experience of autonomy, in which autonomous actions are performed intentionally, with understanding, and without controlling influences (Faden & Beauchamp 1986, 238). They put understanding and freedom from control on continua—they recognize that these factors are not binary, and that individuals can experience degrees of understanding and coercion. Autonomy itself, therefore, exists on a continuum, with these variables interacting with each other. If an action is coerced, there is no degree of intentionality or understanding that can make it autonomous, just as no degree of intentionality or freedom can make an action autonomous if it is not understood.

Faden and Beauchamp developed their model of intentionality in light of both philosophy and psychology—the agent in question must have a concrete plan and act to follow up on it (instead of acting accidentally or on habituated and automatic behaviors). Their picture of psychological understanding is based on propositional reasoning and the degree to which an agent has justified beliefs about what he or she is doing. In order to demonstrate understanding, the moral agent in their model must describe both the intended action and its consequences, taking into consideration that an action may be performed with something less than complete understanding or in the presence of false beliefs. The model of controlling and coercive forces requires a separate understanding of will, voluntary action, and control—they note that an agent may fully intend and will an action even if it is influenced or controlled. An agent who is being manipulated, however, is not exercising autonomy.

As with Veatch, there are elements that are intuitive and appealing—it makes sense for us to understand the role manipulation plays in undermining our ability to act autonomously, and it makes sense to integrate concepts in psychology and philosophy in defining intentionality and understanding. However, as with Veatch, there are also

compelling reasons to argue that this model is still predicated on an unrealistic cognitive agent. The historical defense provided to Veatch loses some of its weight as Faden and Beauchamp's model recognized the need to integrate psychology into autonomy and Tversky and Kahneman's *Judgment Under Uncertainty* had already been published, meaning that knowledge and insight into cognitive heuristics and biases were established enough to put forth a collection of papers for broader consumption. The larger problem, however, is that research has demonstrated a number of potential internal influences which can undermine a rational agent's thought process yet which can still yield an "autonomous" decision per this model.

Beauchamp and Childress

The third noncognitive model under consideration is, by far, the most popular methodology in contemporary medical ethics—the principlism of Tom Beauchamp and James Childress (2012). Currently in its 7th Edition, their *Principles of Biomedical Ethics* has remained highly influential in the field, and students entering clinical practice are instructed in the weighing and balancing of beneficience, nonmaleficence, justice, and autonomy. The system is rightly praised for its blending of deontological and consequentialist methodologies (similar to Veatch's blend of consequentialist and nonconsequentialist approaches), which produces a versatility and applicability in a variety of clinical contexts.

Beauchamp and Childress do not make autonomy lexically prior in their system—they recognize that there may be circumstances in which personal autonomy interests are outweighed by other, more essential claims. However, they do place significant importance on it, maintaining a framework in which autonomy must be respected and requiring significant contextual concerns to value other principles ahead of it. They understand autonomy to involve as a minimum the ability to make one's own decisions intentionally, free from outside control, and from limitations that may prevent one from making meaningful decisions (e.g., a lack of understanding). Respecting autonomy in their model requires us to recognize patients' right to hold views and opinions, the right to make choices, and to act upon their opinions and beliefs. They argue that this respect requires both positive and negative duties from us: obligations to disclose information and foster autonomous decision-making, as well as obligations to avoid imposing constraints on autonomous action. This duty does not extend to patients experiencing diminished autonomy, like immature children, those who are ignorant or cognitively incapacitated, or those who are being coerced or exploited. Thus, our obligations to those

with diminished capacity to make medical decisions can be different from our obligations to an uncompromised patient.

Beauchamp and Childress tie their discussion of autonomy to competence, noting that the defining criteria of the autonomous patient and competent patient are "strikingly similar" despite having distinct meanings (*Ibid*. 116). They argue that we should not adopt global standards of competence (i.e., that we should understand judgments of competence to be task-specific) because there are significant difficulties in the validity and reliability of current tests for incompetence—the "evidence" of incompetence isn't necessarily reliable. Instead, when we are concerned about a patient's competence, we should examine her ability to understand her current circumstances and the information she has received, reason about her life decisions, and formulate a choice or preference. In light of this, they suggest that as the risk of a decision increases (for instance, the risk of death), we can reasonably ask for a greater level of *evidence* supporting a decision, but not a greater level of *competence*.

Beauchamp and Childress are not unaware of psychological issues in decision-making. They are aware of differing levels of understanding, the impact of framing effects, difficulties in processing risks, and other elements that can lead patients into false beliefs, and as a result they argue that clinicians ought to challenge patient perceptions and choices in order to better their autonomy (*Ibid*. 137). They also recognize that there are conditions that can impact the voluntariness of actions, like disease, psychiatric disorders, and drug addictions, which preclude autonomous choice and decision-making. Further, in a discussion of hard and soft paternalism, they recognize that there are cognitive biases and bounded rationality in decision-making, but they argue that these ought *not* to be bases for challenging patient autonomy, as it strays into opaque and potentially abusive hard paternalism (*Ibid*. 219). As such, they are aware of relevant challenges to the notion of a Kantian rational agent. Unfortunately, this poses significant problems for their model.

First, it suggests a contradiction, in that they encourage clinicians to challenge their patients' perceptions and choices when they are predicated on false beliefs based on misunderstanding, framing effects, and risk-processing deficits, but caution against challenging their patients' perceptions and choices when they are predicated on bounded rationality and cognitive biases, despite these factors potentially producing misunderstanding and risk-processing deficits. Second, the recognition of bounded rationality suggests awareness that there are essential limits to conscious reasoning and that there is a body of literature exploring alternative explanations for human cognition, including emotional processing, backstage cognition, dual processing models, etc. Simply

put, it isn't clear how one can argue for an overly rational model of cognition when one is aware of myriad empirical data undermining this position.

The preceding analysis is not meant to fundamentally scuttle the theories discussed. They have individual strengths and weaknesses that ought to inform subsequent models. It makes eminent sense to establish prima facie duties, for instance, and to value a collaborative relationship between physician and patient. It makes eminent sense to recognize that ethics is pluralistic, and that it is unlikely that any single principle ought to carry universal and absolute weight. It makes sense to draw upon a variety of philosophical outlooks in offering justification for action, or in discerning the appropriate moral methodology for a given ethical conflict.

However, it does not make sense to predicate an ethical theory on a model of human thought that does not exist. Fiating cognitive abilities amounts to requiring us not to be human when exploring ethical dilemmas or making treatment decisions. It makes no sense to believe that we exercise control over avolitional backstage processes, or to ignore demonstrable sources of error in decision-making, especially when the choices to be made are potentially the most meaningful and most irrevocable of decisions. It makes no sense to suggest that identifiable sources of error ought not to be eliminated as much as possible, to ensure that the choice made is a genuine reflection of the patient's desires, and is not simply the disease process speaking for them. The models that follow attempt to elicit these sources of error, while reaching fundamentally different conclusions.

Cognitive models of autonomy

In contrast to the homuncular models, the cognitive models explore the backstage and automatic elements of patients making health decisions. Four principle models are examined, and the strengths and shortcomings of each are noted. A recurring theme in these critiques is that cognition is fundamentally influenced by a variety of factors not considered in the homuncular models. As such, by their very nature, they present models of autonomy that have much more empirical and ecological validity—they are autonomy models of actual human beings, rather than of idealized cognitive agents.

The first cognitive model to be considered is that of Redelmeier, Rozin, and Kahneman (1993). Contrary to the homuncular models discussed earlier, they argue that the 'ideal' decision maker—characterized by the agent who gathers all available information, calculates the risks and benefits of every option, and then selects the optimal choice—simply does not exist. Instead, actual decision-makers employ cognitive heuristics to simplify situations and find palatable solutions. Additionally, they are

influenced by a variety of sources, including external and internal stimuli, and can be strongly affected by how information is framed. Minor shifts in decision context, option order, defaults, or semantics can radically alter perception and subsequent processing, and yet these are not necessarily changes of which we are aware.[4] Further, individuals can demonstrate a phenomenon called 'hindsight bias'—when individuals learn of the outcome of a given action, this knowledge affects their assessments of the likelihood of that outcome occurring. This is to say that individuals tend to ignore contradictory evidence, focus only on corroborating evidence, and overestimate the probability of the outcome. This is a significant concern in medical liability cases, for instance—arguments that a clinician "should have seen this coming" demonstrate hindsight bias. In the context of medical treatment, this can affect patients' perceptions of their current situation (e.g., 'it was inevitable that I would get cancer'), and can feed into other sources of cognitive error.

They note that many research studies fail to take into account salient features of the patient experience when exploring outcomes and efficacy. There are emotional aspects of being a patient, for instance, which are reflected in one's sense of well-being and validation. Patients, as a result, often seek medical care for sympathy and reassurance (Redelmeier, Rozin, & Kahneman 1993, 74). This presents a difficulty for research, however, in that these emotional valences and experiences are difficult to quantify in the same way as one could quantify physical or mental disability. Difficulty in measurement, however, does not translate into irrelevancy.

This emotional content complicates medicolegal issues as well. They note that the process of informed consent requires the clinician to disclose the risks, benefits, and outcomes of particular interventions. Ostensibly the patient then decides which option best suits his needs and values, but this concept does not take into account the plasticity of human emotion—his needs and values may not be the same once the intervention

4. This really is a remarkable phenomenon. Environmental cues, for instance, have been demonstrated to be a confounding variable in research, and as such, are controlled as much as possible. Presentation order has been shown to demonstrate that individuals have a tendency to choose the last option presented to them—even if the items presented are identical—and that they will offer fabricated justifications to explain why that particular option was different than the others. The presence of defaults has also been demonstrated to affect cognition—studies have demonstrated that many individuals have a tendency simply to accept default options when presented with a choice. Finally, word choice affects salience—it has been demonstrated that individuals view information differently when it is seen as self-relevant; this perception, however, can be affected by whether the individual properly understands the terminology (e.g., there will be a difference in responses between asking someone if they are diaphoretic versus asking them if they are sweating).

has been selected and performed. They note that "psychologists have shown that people are prone to err when making decisions about long-term consequences because they fail to anticipate how their preferences will change over time" (*Ibid*. 74). This is not limited to medical settings—studies have demonstrated that attempts to forecast how one will feel produce errors in such diverse conditions as being fired from one's job to winning the lottery. We have a tendency to believe erroneously that the joy or sorrow we are experiencing now will continue unabated for the foreseeable future. As a result, they suggest that the informed consent process include an appreciation of changes over time, and that patients might benefit from including "statistics and interviews of people who underwent each therapeutic alternative months of years previously" (*Ibid*. 74). As a corollary to their suggestion, it would seem that in the case of forgoing treatment, comparable information might be included, if available.[5]

A special case is presented for patients who are experiencing a recurrence of their illness—some conditions are long-standing with periods of remission (cancer, for instance, or multiple sclerosis). Initially, one might be more inclined to accede to their wishes, as they have already experienced the positive and negative effects of the given intervention. However, even this first-hand experience is not necessarily accurate. They note that memories can also be inaccurate and subject to error.[6] As such, we should not simply defer to patients' prior experience—they may have a distorted sense of the experience (Redelmeier, Rozin, & Kahneman 1993, 74). In light of all of these concerns, they caution that the process of medical decision-making must involve clinicians providing guidance about medical information, but also about common cognitive errors. This is not, however, to claim that clinicians are in a privileged position—the clinician may employ the same kinds of errors he is seeking to prevent in his patient (Dawson & Arkes 1987).

This model provides a more accurate picture of actual cognitive processing in decision-making, but it is hardly a complete ethical theory. Rather, the article serves as an effort to translate the existing heuristic and biases literature into clinical settings, and to

5. Clearly this may present a problem, as individuals electing to forgo treatment might not necessarily be in any shape to provide said information. Other methods of providing this information might include patient testimonials (written or video), contact with surviving family members, etc. While there are difficulties in securing this information, it is not impossible in any sense of the term.

6. This is not a new claim—Hume, for instance, noted this phenomenon in his epistemology: our (simple and complex sense) impressions cannot be mistaken, but our recollections of those complex sense impressions are fallible. It is quite easy for us to misremember events, locations, and experiences, amplifying certain characteristics and suppressing others. As such, personal recollection and experience are not necessarily infallible guides for action.

make clinicians aware of the issues that they will have to face. More developed theories of autonomy are found in the arguments and models presented next.

Grisso and Appelbaum

Like Beauchamp and Childress, Thomas Grisso and Paul S. Appelbaum (1998) stress that the concepts of autonomy and of competence to consent to (or refuse) treatment are related, arguing that competence to consent necessarily involves four criteria. First, it is necessary that the moral agent be able to express a choice—this is not tied to any particular medium of communication (e.g., the patient does not need to be able to speak to do so), but rather, the patient must possess the ability to make his or her choices known. Second, the patient must be able to understand the information germane to the health care decision. If the patient cannot understand the information at hand, there is no way to act upon it or to voice a preference for one intervention over another. Third, the patient must appreciate the significance of the information and the expected outcomes. If there is no way for the patient to gauge risk or to weigh outcomes, there is no way for the patient to take ownership of the decision—there is a fundamental disconnect between the decision and the outcome. Fourth, the patient must be able to reason with the germane information in a manner that allows him or her to logically weigh treatment options. If a patient cannot reason and deliberate about the decision, there is no manner by which he or she can make a genuinely autonomous choice—it is akin to being asked to write a paper without having any writing implement—some organization may be possible, but clearly the ultimate goal will not be able to be realized. These four criteria are not to be understood as being 'all-or-none' principles—that is to say, each of these criteria exists on a continuum; patients manifest different abilities for each at different times. As such, like Beauchamp and Childress, Grisso and Appelbaum argue that competence is not to be understood globally, but is task specific. Ethical judgments must be cognizant of each of these criteria, but "in practice, not all of them uniformly will be 'required'" (Grisso & Appelbaum 1998, 33). Further, they reject appeals to competence criteria based popular wisdom—i.e., they reject competence criteria tied to whether most people would consider the judgment wise or correct. As such, respect for autonomy in their model requires us to respect patients' decisions despite apparent eccentricity or inadvisability (although cases of gross deficiency to make a choice do not enjoy similar protection). These criteria individually are necessary, but not sufficient, for autonomy—a marked inability to meet one of these criteria would render the autonomy of the decision

suspect, but being able to meet one of these criteria is not sufficient evidence to render the autonomy of the decision beyond reproach.

The most referenced criterion is that of *Understanding*—Grisso and Appelbaum note that courts often rely upon this in decisions about competence (*Ibid*. 38). The concept, however, is quite tricky—the underlying mechanisms and processes of the 'Understanding' construct are not well known or easily defined, involving a list of physiological and psychological processes required to translate an experience into a coherent conscious model of it. This complex series of events is not the only mechanism by which cognition is influenced. There are a host of medical disorders, medications, and other injuries that can profoundly affect cognition. The ease with which disruption occurs facilitates examination and assessment—if a lack of understanding seems evident, there is reason to suspect disrupted underlying cognitive mechanisms. This is not, however, a clearly defined case of cognitive deficiency—they note that patients may *appear* to misunderstand information when the actual underlying mechanism is *miscommunication* (*Ibid*. 41).

Grisso and Appelbaum note that *Appreciation* as a competence standard refers to whether patients appreciate that they have a disorder and acknowledge the consequences of that disorder and its treatments (*Ibid*. 42–43). This use of the term parallels other authorities who refer to an absence of this appreciation and acknowledgement as demonstration of holding objectively false beliefs, explicable in terms of definite cognitive distortions. A caveat is introduced, however, in that this lack of appreciation or acknowledgement must be due to more than disagreement with the diagnosis. They note that several conditions are necessary to demonstrate that a distortion is present, rather than simple disagreement. First, the underlying beliefs the patient holds must be substantially irrational or unrealistic. There is a significant difference between doubting a diagnosis because conflicting information was presented or there is evidence of clinical disagreement and doubting a diagnosis because one believes that he has superhuman powers.[7] Their second criterion is that the belief must be the consequence of impaired

7. A personal anecdote serves as a quick example—a patient experienced a painful swelling on her foot and lower leg following a ballet rehearsal. The first clinician to examine her in the Emergency Department ruled out torn ligaments or tendons, noting that while the swelling had abated, a rash-like discoloration remained. Operating on the premise that it was either a reaction to a bacterial or viral infection of the fascia, he contacted infectious diseases and admitted the patient for what would amount to a ten-day stay. The rash did not respond to the treatments provided, and, in fact, the antibiotics administered provoked a further reaction on the patient's hands and arms. The patient and her family became quite skeptical about the diagnosis, despite the insistence by the clinician that it was an infectious disease. Eventually

cognition or affect. This is necessary in light of the objections of established religions to specific aspects of otherwise routine treatment (e.g., Jehovah's Witnesses prohibitions on using blood products). Some of these systemic beliefs sets may be considered by the clinician to be eccentric, but that does not mean that they can be ignored. Their third criterion is that the belief must be relevant to the patient's treatment decision. If the patient is exhibiting distorted cognition that does not reflect on the treatment decision at hand, it is not germane to an assessment of Appreciation. If a patient maintains the belief that gravity does not apply to him, but manifests no treatment-relevant cognitive distortions, there is no compelling reason to doubt his ability to appreciate other information.[8]

There is a common reaction in medicine that patients are expected to react negatively to bad health news—in fact, many consider it a sign of pathology if bad news does not engender some manner of depressive reaction. However, this can have a profound impact on the course of treatment—clinicians can quite easily endorse decisions of questionable competence, as the depressive symptoms can be masked by the expected grief (Grisso & Appelbaum 1998, 51). In light of this, it may be preferable to err on the side of caution when there is evidence of cognitive distortion. Not all cases will be clear cut, and will likely require significant sensitivity to the biopsychosocial elements of the disease and its pathophysiology.

Their *Reasoning* criterion requires that patients be able to engage in logical cognitive processes using the information they understand and appreciate. As noted above, there is significant concern that one may be given information but not be able to use it. Cases of anterograde amnesia, for instance, present challenges to processing because of the speed with which information is forgotten. Alzheimer's dementia and cerebrovascular accidents near memory structures carry similar risks—they prevent individuals from

---

an orthopedist—a friend of the family—visited, and immediately declared that the mysterious 'rash' was simply a bruise that resulted from torn ligaments; the hospital orthopedist concurred, and the patient was discharged later that day. Clearly the patient's and family's disagreement with the diagnosis was not unreasonable or irrational. Questions about the rationality of the patient's and family's beliefs would have been more appropriately raised had she claimed that she was immune to all diseases and infections.

8. For instance, early in my teaching career, I worked with patients with schizophrenia of a variety of severities and degrees of subsequent cognitive impairments, including auditory and visual hallucinations, perceived conspiracies and threats, irrational degrees of grandiosity, and with varying degrees of insight into their conditions. This has not prevented them from being able to engage and process information in many other areas of their lives, nor has their illness prevented many of them from being able to appreciate their clinical situation and course of treatment.

being able to work with new information presented to them. As such, clinicians assessing competence in patients with conditions similar to these ought to be aware of potential influences. Grisso and Appelbaum caution, however, that this criterion ought not be used to deny individuals their right to autonomy simply because they employ non-normative approaches to information processing. They note that most, if not all, individuals fail to meet idealized standards of decision-making in everyday situations, and that these deficits may become more apparent in times of crisis. As such, they stress that Reasoning deficits should focus on cases in which "a patient's mental abilities are so impaired by illness or disability that even basic functioning with regard to these considerations is seriously and negatively influenced" (*Ibid*. 55).

Grisso and Appelbaum stress that certain cases merit greater attention than others—significant changes in mental functioning (generally with behavioral correlates) should serve as warning signals that cognition has been altered.[9] While refusal of treatment or evaluation may be atypical for a particular patient, that alone does not suffice to demonstrate that cognitive changes have occurred, but it should serve as a warning sign. They note that patients with organic impairments are especially prone to decisional incapacity (e.g., dementias, deliriums, etc.). They further note that while depression has been a frequently studied group, the results have varied, suggesting that the differences in the research findings may reflect different degrees of depression, with correspondingly different degrees of impairment. Further, influencing factors are additive—comorbid psychopathologies can exacerbate cognitive distortions and disabilities, which are further exacerbated by medical illness and pharmacological interventions, with polypharmacy being especially problematic (and, among elderly patients, all too common). Finally, while age itself does not *necessarily* reduce competence, they note that it does increase susceptibility to decisional impairment.

The metaphor proposed by Grisso and Appelbaum is a scale whose cups are labelled 'autonomy' and 'protection'. The fulcrum is off center, allowing autonomy a natural advantage (representing social preference for personal autonomy). In the context of a patient either providing or refusing consent to a particular treatment, assessment of information is added to each side, with evidence supporting competence filling the 'autonomy' cup, and evidence undermining competence filling the 'protection' cup. Clearly in this model it requires more evidence to countermand the patient's autonomy

9. By this they mean patients behaving in manners contrary to their normal presentation and personality (e.g., fastidious patients who have become slovenly, gregarious patients who are withdrawn and asocial, etc.). They note that elderly patients are particularly at risk for manifesting these types of changes.

than it does to countermand the duty to protect him or her. It is very uncommon for a patient to completely lose her capacity for Understanding, Appreciation, or Reasoning—as these are continuum concepts, it is more likely that the patient's abilities will simply experience a reduced capacity. As such, clinicians need to be cognizant of the degree of impairment when balancing the metaphorical scale. The consequence of maintaining this balancing metaphor is a sliding standard of competence dependent upon risk-gain ratio analysis of the intervention in question. The fulcrum of the scale is also subject to adjustment—Grisso and Appelbaum allow the clinician to move the fulcrum dependent upon the treatment preferences of the patient. For instance, if the patient elects a procedure that has a less desirable risk-gain ratio than the intervention proposed by the clinician, the fulcrum may be adjusted slightly, requiring more evidence of competence than would normally be required. The patient, however, would need to be duly informed that greater decisional capacity must be demonstrated before the preferred treatment is initiated.

There are significant strengths in this model—for instance, its awareness of the complex interactions of illness and cognition, its understanding that normal judgment can be biased by a variety of sources not normally accounted for in other autonomy models, etc. There are some concerns, however, in that it does not acknowledge that clinicians themselves can demonstrate cognitive biases. Studies have demonstrated that clinicians can focus on one particular diagnosis and ignore others.[10] The very same cognitive heuristics that plague patient decision-makers are found in the clinical staff treating them; as such, awareness of cognitive biases and distortions is not a one-way process. The model proposed by Grisso and Appelbaum would be strengthened by a more dialogical approach, in which the distortions and biases of both physician and patient are exposed and challenged.

10. I recall a passionate discussion I had with one psychiatrist who insisted that a patient was a chronic paranoid schizophrenic, simply because he had carried that diagnosis for several years. The difficulty, however, was that the differential was wider than this particular diagnosis—specifically, he showed considerable evidence of a frontal lobe syndrome. Specifically, he chronically abused crack cocaine (which in long-term abusers produces feelings of paranoia, as well as auditory, visual, and tactile hallucinations), per his family history he had had a traumatic brain injury prior to the onset of his symptoms, his personality was very childlike, irresponsible, and sexually preoccupied, and his affect was not flattened (flat affect is characteristic of chronic paranoid schizophrenia).

Katz

The psychodynamics of the physician-patient relationship is a key element of the autonomy model proposed by Jay Katz (2002). Katz notes that there are many definitions of autonomy, but chooses to focus on what he refers to as 'psychological autonomy'—the capacity of persons to exercise the right to self-determination, which includes their ability to reflect on the choices they have made. He further notes that current conceptions of autonomy make a significant number of psychological assumptions which go unexplored in the literature. Contemporary medical ethics is dominated by abstractions—specifically, abstract norms that generalize conduct in a manner that is inappropriate when considering how human agents actually behave. Ethicists have a tendency to rely upon the theories of Kant and Mill, among other philosophers, to relate the abstract formal norms to material situations. These abstractions contain implicit models of the human psyche which are not developed or clarified, which is unfortunate, in that "[a] careful scrutiny of many philosophical, moral, political or legal principles reveals all kinds of hidden, albeit woefully mutilated, assumptions about human nature" (Katz 2002, 108).

Paradigmatic in medical ethics are the assumptions made by Immanuel Kant—his idealized moral agent is a being of pure rationality; in the ideal agent, moral decision making will not be influenced by whims, emotions, or personal inclinations. Katz notes that current philosophers have championed this model—but the problem lies in that the model itself is untenable. Kant (1996) himself noted that he was making a distinction between an *idealized* moral agent, which he distinguishes from *actual* moral agents—it was a *theoretical* model, not a *practical* model. Kant's model recognizes only one aspect of human behavior as relevant to moral and ethical decision-making—the capacity for rational thought—but ignores or devalues many other aspects of our behavior, which is contingent upon other processes, some of which are completely irrational. Because we can be influenced by so many different aspects of our rational and irrational nature, Katz notes that Kant's model is simply impractical, and therefore is irrelevant in practical situations.

As a result, Katz adopts an autonomy radically different than Kant's ideal—psychological autonomy. Katz's clarifies his definition of the concept, noting that as an *ideal* definition, "psychological autonomy refers to the capacity of persons to reflect, choose, and act with an awareness of the internal and external influences and reasons that they would wish to accept" (2002, 111). Katz stresses that this is an ideal—the sheer volume of internal and external influences makes it impossible for a moral agent to ever

be *fully* aware of them all.[11] Self-reflection and dialogic interaction with others can help to draw out unconscious influences, returning them to the control of the agent.

Katz notes that past discussion of psychological capacities of moral agents has tended to reflect psychopathology instead of underlying motives, i.e., questions of incompetence. He supports those who conclude that only the choices of clearly incompetent patients should be rejected—he argues that it quite different to recognize the sources influencing a patient and interfering with the patient's choice when one believes that they have made the 'wrong choice.' There are implicit dangers in raising psychological objections to patient autonomy—he notes that exceptions to autonomy can be too readily 'found' and that the purview of psychological objections are too far-reaching and too difficult to control. This represents a significant break between Katz's model and my own—while I can appreciate his concern regarding the ease with which questions and challenges to autonomy can be raised, it would seem that the circumstances and the choices to be made would dictate the standard of psychological evidence necessary to maintain patient autonomy (as per Grisso and Appelbaum's model). I will return to this objection below.

At this point, Katz develops the sense of the unconscious employed in his model. Employing a psychodynamic approach, he breaks from other models which suggest that unconscious elements are to be identified, evaluated, and potentially discarded. Specifically he notes the central role of the unconscious in normal decision-making—the psychodynamic perspective seeks to *understand and account for* unconscious influences, rather than *identifying and eliminating* them, as well as identifying potential conflicts between conscious and unconscious motivations. Further, the conscious/unconscious split is not the only germane factor—cognitive modelling of autonomy must take into account the rational/irrational split, as our decision-making process incorporates both. It is extraordinarily rare to find actions that stem from only one motivational source, and the rational/irrational mixture are idiosyncratic, and vary with the individual's situation. In Katz's model, 'rational' and 'irrational' reflect "capacities for adaptation to the external world, that is, persons' conscious and unconscious efforts to reconcile their internal mental processes with the external possibilities and limitations of the world in which they live. They denote persons' abilities to take reality into account and to give some

11. In discussing internal influences, Katz is arguing from a Freudian perspective on conscious and unconscious processes, instead of the sense of the conscious, unconscious, and preconscious cognition developed here. The two are very different—the unconscious, for instance, is the domain of libidinal urges, mediated by the ego and superego in Freudian thought, while unconscious processes like heuristics and biases, information integration, and automaticity are what is meant by the term in my argument.

account of the conflicts between their inner and outer worlds to themselves and others" (*Ibid*. 117). As a result, ideal decision-making will be a dialogic process, in which the idiosyncrasies of both the patient and the clinician can be explored, leading to a greater understanding of the motivations and thought processes of both. This dialogue is not likely to reveal all unconscious motives, but it can reveal more than might be accessible solely through introspection and reflection.[12]

This model has immediate consequences for individual autonomy and liberty. Katz notes that it immediately undermines two concepts in the autonomy debate—radicalized patient autonomy, and standards of perfect understanding in the clinician. Instead, it calls for great introspection and reflection; freedom requires, in Katz's words, "constant struggle and anguish with oneself and with others" (*Ibid*. 121).

By being aware of the limits of human thought, both conscious and unconscious, rational and irrational, clinicians and patients can achieve a greater understanding and awareness of their own thoughts and motivations, and allow them to recognize how their perspectives and experience have influenced them directly and indirectly. This, in turn, gives rise to greater freedom in decision-making—the more motivational factors we are conscious of, the more control we exercise in the decision-making process. This will never produce absolute control, however, and as such, there is always an influence of unconscious and irrational factors in human thought. As such, Katz argues that the first, necessary step in self-determination is self-reflection and reflection with others. This reflection may not produce agreement with the physician and patient, but it can clear up misunderstandings and misperceptions. He still opens the door to physicians being able to interfere in patient decisions (and hence to weak paternalism in Beauchamp and Childress's sense of the term), but he stresses that neither party is asked to submit to the other, and that conversation and shared decision-making prevent significant harms.

If our aim is to facilitate autonomous decision-making, a recurring theme in multiple theories of medical ethics, it seems that conversation and mutual exploration of motives and thought processes are necessary foundational criteria. But what should be done if the patient insists on medical decisions fundamentally at odds with the opinion of the clinician? Katz argues that if we adopt the psychological autonomy model he proposes, clinicians will be required at times to accede to 'foolish choices'—as a matter of principle

12. This is comparable to the adage that 'two heads are better than one.' Individual perception tends not to be self-challenged; the presence of another individual capable of evaluating both the situation as well as the other individuals perception.

of respect, the clinician does not possess the ability to simply overrule any decision which he feels to be ill-advised—I will address this aspect of Katz's model below.

Katz's allows for clinicians to disobey a patient's choice only when two conditions have been met (*Ibid*. 157–158). First, the consequences of the decision must pose significant risks to the patient's immediate physical condition. Katz clarifies this by limiting it to cases in which the patient's illness has interventions which have a good chance of preventing death or persistent serious injury, and when such outcomes are likely in a relatively short period of time. The second condition requires that the patient's cognitive processes are so seriously impaired that neither the clinician nor the patient can understand each other. If there is no apparent means of overcoming the communication barrier, then it is reasonable to proceed in the patient's best medical interest. These are very limited conditions, to be sure, but Katz argues that one ought to err on the side of autonomy. This does not create absolute patient autonomy, however, as Katz is cognizant of challenges which might arise as a result, and argues that if they are unable to reach an agreement, then the doctor and patient should either work within limits set by the patient or go their separate ways. As such, significant authority remains with the patient, but not total authority—respect is a principle that is not unidirectional. Many theories of medical ethics note that clinicians are not automatons—they have moral values and beliefs, just like the patient. One cannot expect a clinician to ignore her own important principles in medical decision-making.

There are significant strengths in the model proposed by Katz. It is clear that recognition of the complex cognitive processes underlying decision-making is emphasized in this model. As a corollary, recognition that both patients and clinicians carry with them their own sets of rationalities and irrationalities is an important step in shared decision-making. This model explicitly requires the identification and exploration of unconscious cognitive factors for both (or all) parties involved in decision-making, in an effort to increase understanding. This allows for critical insight that might be unavailable were one to attempt simple self-exploration and self-reflection. The emphasis on a dialogic process as a requisite first step towards self-determination clearly demonstrates the need for the patient to understand himself before he can make informed decisions. It is quite clear that we cannot make meaningful decisions if we are unclear as to what it is that we want. We can certainly make choices, but it is evident that they may not actually reflect our values or beliefs—in short, they will lack the 'self' criterion of self-determination.

However, there are some concerns about Katz's model as well. First, it is unclear that one ought to adopt a Freudian model of the unconscious, as there are significant

methodological, empirical, and theoretical concerns about the Freudian model.[13] It is clear that unconscious processes influence cognition, but the empirical data and research support a model of unconscious processing quite different from Freud's theories (e.g., automaticity, heuristics and biases, and emotionally-valenced memory and recall). As such, when unconscious motivations are discussed later, it will not be in the terms Katz's proposes.

Second, the criteria set by Katz for incompetence appear to be too high. It is understandable that he would establish such strict criteria in light of the psychoanalytic model he proposes, which integrates the unconscious, but as that methodology is suspect, it seems reasonable to question the need for such restrictive criteria. This is not to say that clinicians ought to have *carte blanche* in deciding which decisions to accept or to reject, but it certainly suggests that the standards for rejecting bad choices ought to be lowered. It is clear that cognition is dependent on a variety of factors, of which we are only aware of the surface phenomena. It is likewise clear that our cognition can be affected in manners great and small at a variety of levels of reduction. It would therefore seem to be reasonable to suggest that clinicians have more leeway than Katz's proposes in challenging the decision-making process of patients, who by their nature are more vulnerable to influences due to medical illness, pharmacology, and potential psychopathology. I do not challenge the idea that patients have the right to make bad choices; I do challenge the idea that this right is an absolute, especially as the consequences of their decisions increases in severity. As suggested earlier, it seems that a quite compelling case can be made for a sliding scale of autonomy, contingent upon the severity of the predicted outcomes, with the most scrutiny applied to terminal decisions.

13. In terms of *methodological* concerns, Freud was not research-oriented. The case studies he selected were not experiments—they were self-selected case studies designed to develop the theory, not test it. In fact, a recurrent criticism of Freudian models is that they do not translate easily—if at all—into testable variables. There are *empirical* questions as well—Freudian psychotherapy and analysis requires significant time and effort—it is common for patients to see their analyst for years before any insight is drawn. This is clearly beyond the purview of a normal in-patient stay. It is much more likely that Katz is advocating a more superficial variant of Freudian analysis, but even in this abbreviated sense, it remains unclear that the average clinician would have the requisite training or understanding needed to identify unconscious motivations. The *theoretical* concerns raised stem from Freud's own statements—as he approached the end of his life, he raised his own concerns as to whether psychoanalysis was actually helpful. If the founder of the school of thought questions its use, one ought to be skeptical about arguments built from the suspect theory.

Anderson and Lux

Higher cognitive standards are established by Anderson and Lux (2004), who argue that the keystone of autonomy and self-determination is 'accurate self-assessment,' and that autonomy is contingent upon an ability to recognize impairments in one's own cognitive capacities. They offer the clinical case of 'John'—a patient who experienced severe frontal lobe injury, which severed his optic nerves (as a result, he had no perception of light at all). As a result of his accident, John experienced a fascinating cognitive impairment—he was unaware that he was blind. Consequently, he would attempt to navigate his way around as he would were his vision normal, with the result that he would walk into walls, trip over furniture, and found himself in various dangerous situations for one who cannot see. Anderson and Lux argue that his actions ought not to be considered autonomous, not because of his visual impairment, but because of *his cognitive inability to recognize that he had a visual impairment*. This is to say, they argue, "[a]t least with respect to those actions, he was deeply alienated from himself as an agent" (Anderson & Lux 2004, 280). There are a number of types of agnosognosia (being unaware that one is unaware of a deficit)—visual, auditory, etc.—each of which pose the same kind of problem for one's self-concept. Further, there are multiple conditions which produce similar deficits in one's sense of self—V.S. Ramachandran, Oliver Sacks, and others describe neurological conditions in which a patient experiences a disconnect between sense data and association cortices, sense data and perception, perception and association cortices, sense data and emotional valence, etc.[14] Clearly it is possible to meet previously proposed criteria for autonomy and yet experience a profound deficit in self-perception. As such, it makes eminent sense for clinicians to examine self-perception for accuracy before asking patients about treatment preferences—if their self-perception is unrealistic or bizarre, there is reason to believe that decisions made upon these perceptions will also be compromised.

Anderson and Lux draw parallels to the category of 'insight into illness' in establishing their criterion of accurate self-assessment (Anderson & Lux 2004, 280). A variety of conditions manifest decreased insight—there are several psychiatric illnesses

14. Interestingly, Ramachandran describes a procedure that temporarily alleviated post-stroke agnosognosia. Checking for nystagmus involves injecting cold water into the left ear (one of the tests performed in some brain-death protocols). Ramachandran found that individuals with a variant of agnosognosia regained an accurate picture of their physical condition (albeit temporary) following the water treatment. See S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (New York: Quill, 1998) for more information.

in which the patient categorically denies any illness.[15] Inaccurate self-assessment in Anderson and Lux's sense has three criteria. First, the patient must intentionally undertake a given task. Several authors have noted that intentional action is a requisite part of autonomy and self-determination; accidental actions are not intentional, and as such, are not dependent upon an agent's belief about their skill in performing said action. The second criterion is that the agent believes that she will be able to perform the given task as it is intended. That is to say, the agent believes that she possesses the requisite skill and ability to complete the task. The third criterion is that this self-assessment of capacity must be inaccurate. Specifically, the agent objectively must not possess the requisite skill or ability in question. It must be demonstrable that the agent possesses a deficit that she does not believe she has.

When erroneous beliefs are examined, these self-perceptions are not understood in terms of whether they are subjectively reasonable, but rather whether they correspond with the facts of the case. This lack of insight does not translate into global incompetence—like Beauchamp and Childress's competence model, it is a task-specific deficit. As such, we see that clinicians assessing insight must possess an accurate understanding of the degree of skill necessary to complete the task in question—if the evaluator's criteria for normal function are set too high, it is entirely possible that competent individuals will be judged incompetent. This is not the only continuum involved in testing accurate self-assessments—in addition to standards varying with the task, the self-assessment itself is a statement of probability. Further, Anderson and Lux argue that there is no single threshold for accuracy, and hence no threshold for autonomy—for most individuals and for most occasions, a general self-assessment of one's capacities should suffice. They suggest that the cases in which inaccurate self-assessment produces non-autonomous actions will be severe enough as to be immediately recognizable (e.g., stumbling into furniture that one cannot see, but claiming no visual impairment). Some agents are able to recognize that they are experiencing cognitive deficits, and can act to correct them or to incorporate them into their cognitive modeling. They argue that the capacity (and hence the autonomy) of these individuals is still compromised in some degree, but less than it was before (maintaining the continuum

15. For instance, I worked with a patient for several years who maintained vociferously that while he was the son of a famous martial artist, was engaged to/married to/dating a pop starlet (the relationship would change from day to day), was a commander in the Navy, Air Force, and Army, and was designing ships for NASA, all while playing with the band Metallica, he was most assuredly not schizophrenic.

approach to autonomy. They further note that just as individuals with cognitive deficits can overestimate their abilities, so too can they underestimate their abilities.[16]

Anderson and Lux stress that the establishment of non-autonomous actions requires more than simple demonstration that the patient is making poor choices or has some unjustified beliefs. They suggest that autonomy does include the ability to make mistakes. As such, they stress that in utilizing their proposed criteria, it must be clear that the deficit in question is preventing the agent from exercising self-governance—i.e., there must be something inherent in the deficit that prevents autonomy itself. There are several methods by which this may be assessed, and Anderson and Lux focus on two in particular. First, it is possible to explore the causal link between the action and the source of the action—if the action occurs in such a way as to prevent evaluation of the motives behind one's action, then the causal pathway has been disrupted, preventing the agent from taking ownership of the action. This is a key concept, and one which will be revisited later. The second method by which ownership of the action can be disrupted concerns problems in integrating the action with its motivations—the agent cannot make sense of his motives or is alienated from them (i.e., the agent experiences a baffling "Why did I do that?" moment). If the agent cannot understand and reconcile his motivations with his actions, there is reason to believe that they are non-autonomous. Anderson and Lux note that these two concerns demonstrate the need for integrated actions, as well as a means of registering that integration has not occurred—a feedback mechanism, in short. They note that this feedback mechanism "must be constituted in such a way that the unintelligibility surfaces. For to the extent to which one is unable to note the internal tensions, one is without this compass, which is so crucial for guiding one's actions in the manner we dub 'autonomous.' And this is why rigidly inaccurate self-assessments undermine autonomy" (*Ibid*. 284). In short, absent this feedback mechanism, our compass is broken, and we have no way of knowing whether we are moving in the right direction. For all we know, instead of reaching our goal, we could be simply traveling in circles. The primacy of accurate self-assessment carries with it a three-fold advantage: first, it is neutral in regards to competing theories; second, it is more plausibly linked with self-direction in autonomy; and third, it is more empirically supported in clinical neuroscience (*Ibid*. 285).

16. In fact, this was a frequent topic in the individual and group therapy sessions held in the behavioral health hospital in which I worked. We helped our patients understand and develop their physical, occupational, and psychological skill sets and resources.

The aspect of Anderson and Lux's analysis that is most crucial to the argument developed here is that they extend it to cover mental as well as physical incapacities. Factors like automaticity, cognitive heuristics and biases, and emotional valencing occur outside of our awareness, and constitute significant but correctable sources of error and distortion. It would seem that these types of errors dovetail with Anderson and Lux's analysis; it is necessary to note, however, that they focus their analysis on traumatic brain injuries, rather than on phenomena of cognitive psychology. However, as the psychological phenomena in question have physical bases, it seems evident that such considerations as Anderson and Lux propose ought to be extended to them as well.

As with the other cognitive models proposed, there are significant strengths in Anderson and Lux's model. Meaningful self-direction is impossible if one's compass is flawed and there is no way to check it. To the extent that we can become aware of our own cognitive shortcomings, we can correspondingly increase our personal autonomy.

There are weaknesses to be found, however. First, it is unclear how far back or how deeply they are willing to extend their cognitive analysis. The kinds of deficits produced by the conditions Anderson and Lux consider also produce systematic error, since they produce a recurring mistaken belief. It is unclear, however, whether Anderson and Lux intend for their argument to be extended to the automatic and backstage elements discussed in the present argument. If they are unwilling to extend their analysis to these types of cognitive errors, it would seem a rather arbitrary distinction, and the autonomy model proposed would certainly require clarification.

The second weakness is that while the model raises compelling arguments, it does not establish a clear metric for establishing non-autonomous actions. They do specify some criteria, but they also place these criteria upon continua, which allows for significant room for interpretation. For the autonomy standard to be meaningful, it would seem that a little more structure or clarity is needed for clinical application beyond claims that distortions and corresponding non-autonomy will be immediately recognizable.

A third concern is that this is not a fully-developed theory of autonomy. To be fair, it does not seem to be intended as such, but the criterion of accuracy in self-perception is a necessary, but not sufficient, element of autonomy. It is quite clear that individuals can act in non-autonomous ways while maintaining accurate perceptions of their abilities. Additional criteria, as have been explicated in the previously discussed models, are critical to an accurate and meaningful picture of autonomy.

Conclusion

The model that emerges from this discussion must necessarily take into account multiple factors drawn from the strengths of the homuncular and cognitive models of autonomy. Four key categories of autonomy criteria emerge—foundational, medical, psychiatric, and psychosocial. Each of these categories is necessary for an autonomous action, but none are sufficient. Each will be explored in turn.

Before presenting them, however, there are several caveats. First, it must be made clear that this model ought only to be considered applicable to end-of-life decisions. It is quite clear that this kind of decisional process has little day-to-day validity—the elements discussed are not part of everyday decision-making. However, as has been suggested earlier, a compelling argument can be raised that as the consequences of our decisions become more severe, greater evidence is needed that the action is autonomous. In terminal decisions, it is unclear why a lower evidentiary standard should be preferred. Second, this model is intended for use in cases when a patient is awake, aware, and able to voice her own preferences. Last, quite obviously this should not be understood as a fully developed theory of medical ethics, nor should it be seen as anything other than criteria necessary for autonomous action as evidenced by the theoretical and empirical challenges raised to the autonomy models found in contemporary theories. It is quite possible to incorporate this understanding of autonomy in existing models (e.g., substituting a cognitive model of patient autonomy would not fundamentally undermine Beauchamp and Childress's principlism), albeit in some more than others (this model *does* present a fundamental challenge to models giving disproportionate weight to autonomy, e.g., Veatch).

Medical Criteria of Autonomy

Medical criteria concern issues that are the traditional purview of medical treatment; i.e., these are routine elements that recur in many theories of medical ethics, and are the least likely to cause concern and controversy. There are two key medical criteria for patient autonomy: the absence of a medical condition which directly affects cognition to the point of incapacity (which I will refer to as Structural Integrity), and access to the information typically required for informed consent. Both of these criteria are continuum-based, as disease processes result in different degrees of impairment, and some pieces of information might be more relevant or available than others.

Structural Integrity

The most significant challenge to patient autonomy in the models discussed is a physical impairment which prevents the patient from taking in information or processing it. Dementia, delirium, traumatic brain injury, cerebrovascular accidents, etc., can exert profound effects on the ability of the patient to take in new information, make their preferences known, form associations between concepts or words, etc., all of which are necessary elements of cognition. Clearly any illness which fundamentally disrupts this process prevents the patient from making a meaningful decision. However, because the effects of these illnesses are not uniform, it would be inappropriate to make blanket statements about the degree to which subsequent actions are autonomous or non-autonomous. As such, a threshold point would need to be established, which could employ any of a number of psychiatric and neurological tests (e.g., the Mini Mental Status Exam).

Informed Consent (or Refusal)

The standard protocol for medical intervention involves securing the informed consent of the patient. While the standards of this vary from state to state (e.g., whether the 'batting average'—the clinicians success rate with the suggested treatment—is required disclosure), there is enough commonality to require that the patient be provided with information concerning the nature and purpose of the intervention, alternative interventions (including non-intervention) and their outcomes, risks, probable outcomes of the intervention proposed, etc. This information should be presented in normal language, and should not require the patient to have extraordinary education to understand it. State standards of informed consent could suffice for threshold points (and due to variance, this criterion exists along a continuum).

Foundational Criteria of Autonomy

Foundational criteria of autonomy refer to underlying psychological structures of the decision-making process. Foundational structures are primary and fundamental—absent these criteria, significant doubt can be raised about the autonomy of the patient's decision. There are five criteria in this category: the ability to consider, make, and make known one's preferences (which I will refer to as capacity for preference); intentionality in action; accurate self-assessment; awareness of common sources of cognitive error (which I will refer to as bias vigilance); and dialogue aimed at self-discovery, which includes the willingness to participate in dialogue. There is no lexical priority for these criteria,

and they fit into both absolute and continuum scales.[17] Each of these requires further exploration and clarification.

Capacity for Preference

In this criterion, the moral agent engages in reflection upon the treatment options open to her, weighs their strengths and weaknesses as she understands them, and makes her preferences known in some manner to the clinician (ideally through a contemporaneous statement). By its very nature, this will post challenges, as the interpretation the patient gives to the treatment option will be contingent upon her perception and understanding, which may require further discussion and dialogue with the clinician, to ensure as much accuracy as possible. This capacity for preference is not absolute, in that patients will differ in both the degree of their preferences as well as their ability to communicate them. Patients unable to weigh information or express preferences due to cognitive impairment or illness ought not to be considered autonomous agents, and treating clinicians should defer to a best-interest standard until the impairment is resolved or a proxy decision-maker is identified.

Intentionality

Several theories have noted the necessity of this criterion. For an action to be personally meaningful and autonomous, it must be intended and not accidental or reflexive. It is entirely possible to act without meaning to act, and a number of neurological and psychiatric conditions have demonstrated that involuntary actions can be physical or verbal. As has been discussed above, mental actions are also driven by automaticity, and therefore the agent may find herself acting or thinking in a manner she does not desire. Following earlier theories, this is an absolute scale—either one intends to act or one does not, and it is quite possible to discern between the two. Unintended actions ought not be considered autonomous.

17. As a necessary caveat and matter of clinical significance—I realize that these proposed standards are theoretical, and may have some difficulty translating well into clinical settings (e.g., discussions of backstage cognition). This is a barrier faced by cognitive therapies in psychology, as well—the theoretical concepts will be dependent upon the underlying cognitive capacity of the patient in question. This can be resolved by using age-, understanding-, or education-appropriate terms (e.g., switching "People frequently make systematic cognitive errors in information processing." with "Sometimes we can get so used to thinking about things some way that we forget there are other ways to see it.")

Accurate Self-Assessment

Following Anderson and Lux's argument, agents must have insight into their illness. If a patient demonstrates agnosognosia, whether correctable or resistant, their autonomy has been weakened. If a patient demonstrates a consistent source of error germane to her medical decision-making process, she cannot process the information necessary to make the judgment (or can only do so in a diminished capacity), and as such lack the insight necessary to be self-directing. This analysis extends not just to awareness of physical injury, but also to persistent cognitive errors and distortions. This criterion exists along a continuum, with autonomy increasing as the degree of accurate self-assessment increases.

Bias Vigilance

Given that cognitive biases and sources of error are so prevalent in 'normal' cognition, and that special circumstances may exist in patients with depression, patients must be educated regarding common sources of cognitive error. This does not mean that the patient must hold a doctorate in psychology, but she must be made aware of the ways in which we frequently misinterpret information, emotional information, and memory. This is a continuum criteria, as patient understanding is variable. If a patient demonstrates an inability to understand backstage cognition (i.e., an inability to recognize that thought can be influenced by other conditions [environmental triggers, personal biases, heuristics, etc.]), there is reason to question her autonomy.[18] This criterion ties in directly with Dialogic Self-Discovery.

Dialogic Self-Discovery

As has been demonstrated earlier, it is quite common that we are unaware of the idiosyncratic and systematic slants we place upon the information we take in, or upon the memories we selectively recall. These biases and slants can be explored in a shared decision-making model as proposed by Katz. While the content is somewhat different than Katz's model, in that the clinician and patient are not attempting to explore the Freudian unconscious, the aim is similar—dialogic interaction can provide illumination on those processes that evade self-exploration and reflection. This criterion exists along

18. This argument will no doubt raise significant questions, and so I feel it requires further clarification. I am not arguing that if the patient is *skeptical* about the information they are not autonomous—simple examples can demonstrate heuristical thinking, which should permit the patient to at least be willing to entertain the idea, in an effort to facilitate Dialogic Self-Discovery. If a patient demonstrates a profound *inability* to conceptualize backstage cognition, there is reason to suspect compromised autonomy.

a continuum for two reasons: first, patients will have varying degrees of insight, so the amount of benefit from dialogic interaction will vary from patient to patient; and second, patients will have varying degrees of willingness to participate in dialogic self-discovery. The more open a patient is to self-discovery, the greater the likelihood of an autonomous action resulting. If a patient categorically refuses to engage in dialogic self-discovery, there is reason to suspect compromised autonomy, but not necessarily proof.[19]

Psychiatric Criteria of Autonomy

There is only one principle psychiatric criterion of autonomy: the minimization of any psychiatric comorbidity (which I will refer to as psychiatric minimization).

Psychiatric Minimization

Given the documented underdiagnosis of depression and other depressive disorders in common medical illnesses, given the effect of depression on morbidity and mortality, and given the influence depressive disorders can exert on a patient's cognitive process, it is important to identify and account for any psychiatric comorbidities, and to attempt to minimize their effect on the patient's thought process. This may employ a trial period on an anti-depressant or mood stabilizing medication, cognitive therapy or another talk-based intervention, etc., in an effort to isolate and control thought processes stemming from a depressive disorder instead of the patient's own expressed values. This criterion exists along a continuum, as the severity of depressive disorders varies. This criterion is linked with the psychosocial criterion of authenticity.

Psychosocial Criteria of Autonomy

Psychosocial criteria of autonomy refer to the relational individual—i.e., it recognizes that the individual exists as part of a network of relationships which can exert influences—as well as referring to the narrative individual—i.e., the individual as she exists over time. There are two essential psychosocial criteria: the minimization

19. There is also the possibility that the patient simply does not want to discuss the matter any further for a variety of reasons (e.g., irritation with the clinical staff, fatigue, pain, personality disorder, desire for privacy, guilt, crisis of faith, etc.). In the event that a patient expresses unwillingness to engage in dialogic self-discovery, it would behoove the clinical staff to identify and document the reasons for refusal, alleviate whatever conditions are immediately preventative (e.g., fatigue or pain), and attempt at a later time, when the patient may be more receptive. Reluctance or refusal are not necessarily indications of compromised autonomy.

of external coercion (which I refer to as coercive minimization) and the ownership and congruence of the individual's choices (authenticity). Both of these criteria are based on continua—recognizing that coercion and authenticity are not all-or-none principles.

Coercive Minimization

Moral agents do not exist in a vacuum—even the choice to forgo medical treatment involves at least two people (physician and patient). As such, it makes no sense to fiat a model of radical individualism, as there is significant empirical refutation of this idea. The choices that we make in life affect other individuals in a variety of ways, some strongly and others weakly. This is not unidirectional, however—the relationships in which we engage, personal and professional, influence how we approach problems and decisions. Some relationships can exert significant influence—our motives can shift from egoistic to altruistic, focusing more on how a decision affects someone else than how it affects ourselves. Further, our decisions can be manipulated by others, through bad information and deception, emotional appeals and threats, etc. Most systems of medical ethics reject such manipulations as fundamentally undermining autonomy, a position advocated here as well. This is not to attempt to argue for radical individualism, as this seems to be untenable. However, it does seem plausible that a proper accounting of personal autonomy should attempt to minimize the coercion applied to any individual—it is unlikely that *all* forms of coercion can be accounted for and prevented, but in a decision as serious as the choice to forgo medical treatment—a terminal decision—it seems clear that one would seek to minimize any *undue* influence.

Authenticity

The authenticity criterion is complicated—on the one hand, it is intuitively reasonable to desire for decisions to reflect the values and choices an individual has taken to be her own; on the other hand, humans have the capacity to change, and that inherent plasticity makes it difficult to insist that the individual act in accordance with the same principles at every point in his or her life (e.g., changing faiths from Roman Catholicism to agnosticism, or vice versa). A compromise position would seem to have individuals explore their contemporaneous values, in light of the other cognitive criteria, and in a dialogic process, in an effort to establish which principles should be considered authentic. The individual's decision could then be examined in light of the congruence between contemporaneous, reflected values and the decision made, with incongruence suggestive of compromised autonomy.

The autonomy model proposed above is no doubt open to criticism, as some claims (e.g., authenticity) have been controversial in the literature. However, they are reasonable criteria, when examined in light of the homuncular and cognitive models of autonomy discussed earlier—there is a compelling reason for each element, and the absence of any of them raises fundamental questions as to the autonomy of the action in question.

Psychology and neuroscience have demonstrated that consciousness, our day-to-day perception, our sense of self and identity, judgment, emotions, and intuitions are all predicated upon a number of causal cognitive elements that are outside our awareness—the bulk of our cognition is deterministic and preconscious. This determinism opens up avenues of undue influence into processes we normally assume to be under our control—it should be clear that this assumption is mistaken at best, inhuman and pernicious at worst. We should not abandon ourselves to blind determinism, however—we possess the ability to reflect upon our motivations, and to engage in dialogic interaction with others, who may bring aspects of ourselves to the fore which would remain otherwise inaccessible. As a result, we can take back a measure of control, but only if we engage in honest dialectic and dialogue with others.

In the context of patient autonomy and decision-making, the necessity of this dialogical process is especially evident—patients are already physically compromised, potentially in ways that can exert conscious and unconscious influence over their decision-making processes, above and beyond the normal potential sources of error found in heuristics and biases. Clinicians should be alert for such influences, recognizing that a medical illness can easily mask a deeper psychopathology. Affective disorders are very common, occur more in patients than in the general population, and tend to go unrecognized or dismissed as a normal reaction to their illness. The effect of these disorders, however, is quite pernicious. They fundamentally affect the efficacy of therapeutic interventions, morbidity and mortality, and rate of recovery—ignoring, dismissing, or failing to identify a comorbidity compromises the treatment of the obvious illness. By only treating the surface pathology, we potentially ignore the deeper wound.

Many contemporary models of autonomy suffer from similar shortcomings—while ethics seeks to inform itself of philosophical, legal, theological, and medical constructs, it all too easily ignores the psychological, an unfortunate irony in light of the fundamental connection between cognitive and clinical psychology and ethical ideals of autonomous choice. Ethical theories that dismiss or fail to address psychological constructs are groundless; models derived from inhuman absolutes are so much fancy and fiction. What good is it to describe models of cognition that have little resemblance to how we actually think?

The present autonomy model suggests that decision-making is a complex construct necessarily containing rational and emotional elements, intuitive judgments, and, as a result, potential sources of error. This seems to gel with day-to-day experience—many decisions are made by gut instinct and intuition, instead of a Cartesian rational process methodically and algorithmically exploring all possible influences, outcomes, and variables. This deterministic model gels with the phenomenon of basing day-to-day decisions upon distal causes—early education and environment, role models, learned behaviors, etc. This model suggests that as the severity of the outcomes increases to terminal, increasing reflection upon the causes and motivations of the decision is required—that a genuinely autonomous choice will explore the agent's motivations, identifying and judging the appropriateness of each influence, determining if it is congruent with the value system adopted by the agent as a whole. Decisions stemming from inauthentic elements of the self fundamentally are not expressions of autonomy; if a patient is forgoing treatment, whether to avoid suffering or actively to choose death, we would be remiss not to ensure that it is *her*, and not *her pathology* making the choice. Anything less would surrender autonomy to expediency, would surrender authenticity to apathy, and would surrender insight to obfuscation. The capacity for self-reflection appears to be a defining characteristic of being human—we would do well to use it when we face terminal choices.

References

Anderson, Joel, and Warren Lux. 2004. "Knowing your own strength: accurate self-assessment as a requirement for personal autonomy." *Philosophy, Psychiatry, and Psychology* 11 (4): 279–94.

Ashcraft, Mark H. 1994. *Human Memory and Cognition.* New York: HarperCollins College Publishers.

Bargh, John A. 1989. "Conditional Automaticity: Varieties of Automatic Influence in Social Perception and Cognition." In *Unintended Thought*, edited by James S. Uleman and John A. Bargh, 3–51. New York: Guilford Press.

Bargh, John A. 1997. "The Automaticity of Everyday Life." In *The Automaticity of Everyday Life*, 1–61. Mahwah: Lawrence Erlbaum Associates.

Bargh, John A., and Melissa J. Ferguson. 2000. "Beyond Behaviorism: On the Automaticity of Higher Mental Processes." *Psychological Bulletin* 126 (6): 925–45.

Baumeister, Roy E., and Kristin L. Sommer. 1997. "Conscious, Free Choice, and Automaticity." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 75–81. Mahwah: Lawrence Erlbaum Associates.

Beauchamp, Tom L., and James F. Childress. 2001. *Principles of Biomedical Ethics*. 5th. New York: Oxford University Press.

—. 2012. *Principles of Biomedical Ethics*. 7th. New York: Oxford University Press.

Berkowitz, Leonard. 1997. "Some Thoughts Extending Bargh's Argument." In *The Automaticity of Everyday Life*, 83–94. Mahwah: Lawrence Erlbaum Associates.

Butkus, Matthew Allen. 2006. "Depression, Volition, and Death: The Effects of Depressive Disorders on the Autonomous Choice to Forgo Medical Treatment." Pittsburgh, PA: Duquesne University. doi:10.13140/2.1.3236.9284.

Carver, Charles S. 1997. "Associations to Automaticity." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 95–103. Mahwah: Lawrence Erlbaum Associates.

Chapman, Gretchen B., and Eric J. Johnson. 2002. "Incorporating the Irrelevant: Anchors in Judgments of Belief and Virtue." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 120–38. New York: Cambridge University Press.

Chase, Valerie M., Ralph Hertwig, and Gerd Gigerenzer. 1998. "Visions of Rationality." *Trends in Cognitive Science* 2 (6): 206–14.

Clore, Gerald, and Timothy Ketelaar. 1997. "Minding our Emotions: On the Role of Automatic, Unconscious Affect." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 105–20. Mahwah: Lawrence Erlbaum Associates.

Conly, Sarah. 2013. *Against Autonomy: Justifying Coercive Paternalism*. Cambridge: Cambridge University press.

Dawson, Neal V., and Hal R. Arkes. 1987. "Systematic Errors in Medical Decision Making: Judgment Limitations." *Journal of General Internal Medicine* 2: 183–187.

Donchin, Anne. 2001. "Understanding autonomy relationally: Toward a reconfiguration of bioethical principles." *Journal of Medicine and Philosophy* 26 (4): 365–86.

Eddy, David M. 1982. "Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 249–67. New York: Cambridge University Press.

Einhorn, Hillel J. 1982. "Learning from Experience and Suboptimal Rules in Decision-Making." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 268–83. New York: Cambridge University Press.

Evans, Mark D., and Steven D. Hollon. 1988. "Patterns of personal and causal inference: implications for the cognitive therapy of depression." In *Cognitive Processes in Depression*, edited by Lauren B. Alloy, 344–77. New York: Guilford Press.

Faden, Ruth R., and Tom L. Beauchamp. 1986. *A History and Theory of Informed Consent.* New York: Oxford University Press.

Fauconnier, Gilles. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language.* New York: Cambridge University Press.

Fauconnier, Gilles, and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* New York: Basic Books.

Gawande, Atul. 2002. *Complications: A Surgeon's Notes on an Imperfect Science.* New York: Picador.

Gigerenzer, Gerd. 1996. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103 (3): 592–596.

Gigerenzer, Gerd, Jean Czerlinski, and Laura Martignon. 2002. "How Good Are Fast and Frugal Heuristics?" In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 559–81. New York: Cambridge.

Gilbert, Daniel T., Elizabeth C. Pinel, Timothy D. Wilson, Stephen J. Blumberg, and Thalia P. Wheatley. 2002. "Durability Bias in Affective Forecasting." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 292–312. New York: Cambridge University Press.

Gilovich, Thomas, and Dale Griffin. 2002. "Introdution—heuristics and biases: then and now." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 292–312. New York: Cambridge University Press.

Griffin, Dale, and Amos Tversky. 2002. "The Weighing of Evidence and the Determinants of Confidence." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 230–49. New York: Cambridge University Press.

Grisso, Thomas, and Paul S. Appelbaum. 1998. *Assessing Competence to Consent to Treatment: A Guide for Physicians and Other Health Professionals.* New York: Oxford University Press.

Hájíček, P. 2009. "Free will as relative freedom with a conscious component." *Consciousness and Cognition* 18: 103–109.

Homan, Richard W. 2003. "Autonomy reconfigured: incorporating the role of the unconscious." *Perspectives in Biology and Medicine* 46 (1): 96–108.

Isen, Alice M., and Gregory Andrade Diamond. 1989. "Affect and Automaticity." In *Unintended Thought*, 124–52. New York: Guilford Press.

Jennings, Bruce. 1998. "Autonomy and difference: The travels of liberalism in bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 258–69. Upper Saddle River: Prentice Hall.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

Kant, Immanuel. 1996. *Critique of Pure Reason*. Translated by Werner S. Pluhar. Indianapolis: Hackett.

Katz, Jay. 2002. *The Silent World of Doctor and Patient.* Baltimore: The Johns Hopkins University Press.

Lakoff, George, and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* New York: Basic Books.

Lane, Robert E. 2000. "Moral blame and causal explanation." *Journal of Applied Philosophy* 17 (1): 45–58.

Levy, Daniel A. 2003. "Neural holism and free will." *Philosophical Psychology* 16 (2): 205–228.

Light, Donald W., and Glenn McGee. 1998. "On the embeddedness of bioethics." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 1–15. Upper Saddle River: Prentice Hall.

Logan, Gordon D. 1989. "Automaticity and Cognitive Control." In *Unintended Thought*, edited by James S. UIeman and John A. Bargh, 52–74. New York: Guilford Press.

Meynan, Gerben. 2010. "Free will and mental disorder: Exploring the relationship." *Theoretical Medicine and Bioethics* 31: 429–443.

Miller, Dale T., and Marlene M. Moretti. 1988. "The causal attributions of depressives: self-serving or self-disserving?" In *Cognitive Processes in Depression*, edited by Lauren B. Alloy, 266–88. New York: Guilford Press.

Mischel, Walter. 1997. "Was the Cognitive Revolution Just a Detour on the Road to Behaviorism?" In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 181–186. Mahwah: Lawrence Erlbaum Associates.

Müller, Sabine, and Henrik Walter. 2010. "Reviewing Autonomy: Implications of the Neurosciences and the Free Will Debate for the Princple of Respect for the Patient's Autonomy." *Cambridge Quarterly of Healthcare Ethics* 19: 205–217.

Nisbett, Richard E., Eugene Borgida, Rick Crandall, and Harvey Reed. 1982. "Popular Induction: Information is not Necessarily Informative." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 101–16. New York: Cambridge University Press.

Parks, Jennifer. 1998. "A contextualized approach to patient autonomy within the therapeutic relationship." *Journal of Medical Humanities* 19 (4): 299–311.

Redelmeier, Donald A., Paul Rozin, and Daniel Kahneman. 1993. "Understanding patients' decision: cognitive and emotional perspectives." *Journal of the American Medical Association* 270 (1): 72–76.

Roessler, Beate. 2002. "Problems with autonomy." *Hypatia* 17 (4): 143–62.

Schwarz, Norbert. 2002. "Feelings as Information: Moods Influence Judgments and Processing Strategies." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 534–547. New York: Cambridge University Press.

Schwarz, Norbert, and Leigh Ann Vaughn. 2002. "The Availability Heuristic Revisited: Ease of Recall and Content of Recall as Distinct Sources of Information." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 103–19. New York: Cambridge University Press.

Shepherd, Joshua. 2012. "Free will and consciousness: Experimental studies." *Consciousness and Cognition* 21: 915–927.

Simon, Herbert A. 1990. "Alternative Visions of Rationality." In *Rationality in Action: Contemporary Approaches*, edited by Paul K. Moser, 189–204. New York: Cambridge University Press.

Sloman, Steven A. 2002. "Two Systems of Reasoning." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 379–96. New York: Cambridge University Press.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. 2002. "The Affect Heuristic." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 395–420. New York: Cambridge University Press.

Smith, Eliot R. 1997. "Preconscious Automaticity in a Modular Connectionist System." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 187–202. Mahwah: Lawrence Erlbaum Associates.

Stanovich, Keith E., and Richard West. 2002. "Individual Differences in Reasoning: Implications for the Rationality Debate." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman, 421–40. New York: Cambridge.

Tait, Rosemary, and Roxanne Cohen Silver. 1989. "Coming to Terms with Major Negative Life Events." In *Unintended Thought*, edited by James Uleman and John A. Bargh, 351–82. New York: Guilford Press.

Tauber, Alfred I. 2003. "Sick autonomy." *Perspectives in Biology and Medicine* 46 (4): 484–95.

Turner, Mark. 2000. "Backstage Cognition in Reason and Choice." In *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*, 264–86. New York: Cambridge University Press.

Tversky, Amos, and Daniel Kahneman. 1982. "Availability: A Heuristic for Judging Frequency and Probability." In *Judgment Under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 163–178. New York: Cambridge University Press.

Veatch, Robert M. 1981. *A Theory of Medical Ethics.* New York: Basic Books.

Weinstein, Neil D. 2003. "Exploring the links between risk perceptions and preventive health behavior." In *Social Psychological Foundations of Health and Illness*, edited by J. Suls and K. A. Wallston, 22–53. Malden: Blackwell Publishing.

Wolpe, Paul R. 1998. "The triumph of autonomy in American bioethics: a sociological view." In *Bioethics and Society: Constructing the Ethical Enterprise*, edited by Raymond DeVries and Janardan Subedi, 38–59. Upper Saddle River: Prentice Hall.

Journal of Cognition and Neuroethics

Rolling Back the Luck Problem for Libertarianism

Zac Cogley
Northern Michigan University

Biography
Zac Cogley is Associate Professor of Philosophy at Northern Michigan University. His research interests include moral responsibility, emotion, and other related concepts. His publications include pieces on trust, the blaming emotions, the nature of angry virtue and vice, and the grounds for deserving resentment.

Acknowledgments
I am grateful to feedback from a friendly audience at the Free Will conference sponsored by the Center for Cognition and Neuroethics, a less-friendly PACT audience, and Andrea Scarpino.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Cogley, Zac. 2015. "Rolling Back the Luck Problem for Libertarianism." *Journal of Cognition and Neuroethics* 3 (1): 121–137.

Rolling Back the Luck Problem for Libertarianism

Zac Cogley

Abstract

I here sketch a reply to Peter van Inwagen's Rollback Argument, which suggests that libertarian accounts of free agency are beset by problems involving luck. Van Inwagen imagines an indeterministic agent whose universe is repeatedly 'rolled back' by God to the time of her choice. Since the agent's choice is indeterministic, her choices are sometimes different in the imaginary rollback scenarios. I show that although this is true, this need not impair her control over what she does. I develop an account of when and why the fact that an agent would choose differently impairs control, which provides a novel response to the Rollback Argument.

Keywords

Libertarianism, Luck, Peter van Inwagen, free will, Robert Kane

1. Introduction

Libertarians believe that free-will and moral responsibility are incompatible with determinism. They hold that only causal indeterminism (of the relevant sort) ensures that when an agent acts she chooses between a plurality of options so that, whatever she chooses to do, she was free to choose something else. On the libertarian view, this is required in order for an agent to be morally responsible for her actions. One of the most pressing objections to contemporary libertarian accounts of free-will is luck. Peter van Inwagen's Rollback Argument (2000) has recently gained favor as a way of highlighting the luck problem for libertarians. In this paper, I sketch a response to the Rollback Argument on behalf of libertarians. I argue that the phenomenon of rollback need not be problematic for libertarians. Whether an agent's freedom and control is seen as impaired when we consider rollback scenarios depends on the underlying core agential features of a person. Bringing these agential features to the fore demonstrates that rather than presenting a threat to libertarianism, rollback may actually be a helpful tool for libertarians in further developing theories of morally responsible action.

2. A Libertarian Sketch

According to libertarianism, we are at least sometimes able to make choices that are free and for which we are morally responsible.[1] Libertarians see this ability as grounded in the fact that at least some of our choices are not determined. I will assume an event-causal libertarian view, which explains a person's making a choice by appealing to certain agent-involving mental events that cause that choice.[2] Call the relevant mental events that cause a person's choices her *mental set*. A person's mental set is the collection of beliefs, desires, preferences, intentions, judgments, resolutions, and so on that plays a relevant role in making it the case that she makes a particular choice. If a choice were *determined* that would mean, given a person's mental set at the moment she was choosing, there would be only one choice she could make. So, for example, if my choice to work on this paper instead of go for a run were determined, it would only be possible for me to choose to work on the paper. Given my mental set at the time of my choice, I would not and could not choose to go for a run.

An animating idea of libertarianism, as I understand it, is that if I could only choose to work on the paper, I would seem to lack sufficient control over my choice for it to be truly free. If it were not truly free, I would not be morally responsible for making it. For a libertarian, an agent has sufficient control to be free and morally responsible only when it is true that, given her mental set, she really could choose in more than one way. So in the case above, for me to be morally responsible for choosing to work on the paper it would also have to have been possible for me to choose to go for a run (or choose something else, instead). Thus, my *choice* must not be determined by my mental set in order for it to be free and for me to be morally responsible for it.[3]

Take an example: suppose Anne, a businesswoman in a hurry on the way to an important meeting, must decide whether to intervene in an assault (Kane 2007, 26). Anne's presence at the meeting could aid her career, but she is also sensitive to the help

1. Some philosophers allow for the separation of freedom and moral responsibility. I don't in this paper. I here treat free choice as the control condition on being morally responsible.

2. Libertarianism comes in many 'flavors'; important types include noncausal, event-causal, deliberative, and agent-causal accounts. While I focus on an event-causal picture, I don't imply that invoking other kinds of libertarianism can't help with the luck problem. Perhaps they can; I am merely concentrating on the event-causal picture here. For more on the distinctions between libertarian views and the plausibility of libertarian accounts, see (Clarke 2003).

3. While many define determinism as the claim that given the past and given the laws there is only one possible future, what matters most is that given the person's past mental set and the laws there is only one possible future choice.

needed by the assault victim. If Anne's choice is not casually determined by her mental set, then until she makes the decision there is at least some chance she will choose either to intervene or choose to continue on to the meeting. This means that there are possible worlds just like the actual one where *everything*—including her mental set—is the same right up until the point of the decision, but Anne chooses to go on to the meeting rather than help the assault victim. And there are possible worlds just like the actual one where everything is the same right up until her decision, but Anne chooses to intervene in the assault rather than go to the meeting.

Robert Kane terms this ability to choose either option *plural voluntary control*. According to Kane,

> To have such control over a set of options at a given time is to be able to *bring about* any of the options (to go more-than-one-way) *at will or voluntarily* at the time. That is to say, it is to be able to do *whatever you will* (or most want) to do among a set of options, *whenever you will to do it*, for the reasons you will to do it, and in such manner that neither your doing it nor willing to do it was coerced or compelled. (Kane 1996, 111)

When an agent has such control over her action it is free because she and she alone controls it in the sense that whatever choice she makes is one willed by her. Either way Anne chooses, her choice will be made for reasons (Kane 2007, 29). Let us suppose, then, that Anne in fact chooses to stop and intervene in the assault. She aids the victim in driving off the attacker. Her choice is grounded in her sympathy to the victim's plight and her desire to not let the attacker successfully harm an innocent person. Since her stopping to help is free, libertarians believe she is morally responsible for the choice.

3. Luck

The luck problem results from the fact that Anne's choice to stop the assault—even if well-intentioned—may appear to be a matter of luck. Why? Well, consider that even though Anne stopped to help we know that she was not determined to do so. Importantly, her mental set did not determine that she would stop to aid the victim. It was possible that, given her mental set and the laws of nature, she would have chosen to continue on to her business meeting. It is possible Anne would have selfishly passed by the assault victim.

Christopher Franklin presents a schematic account of the problem that luck presents for libertarianism (Franklin 2011, 201):

(i) If an action is undetermined, then it is a matter of luck.
(ii) If an action is a matter of luck, then it is not free.

If this argument is successful, it establishes that undetermined actions are not free and that agents cannot be morally responsible for performing them. Even worse, it suggests that libertarianism is incoherent. It implies that undetermined actions are free *and* that undetermined actions are also not free! But the argument can be challenged—it all depends on what we take luck to involve. For example, suppose we say that an outcome, action, or choice is 'lucky' just in case it is determined at least in part by *something other than an agent*. For example, my hitting a hole in one was lucky because after I hit the golf ball the wind blew in just the right way so that the ball went into the cup. If the wind hadn't blown in the way it did, I wouldn't have made the shot.

However, if we examine Anne's choice using *this* account of luck, we can see that Anne's choice isn't lucky. Anne's choice to go to the meeting or help the assault victim is undetermined. But the fact that it is undetermined does not imply that it is determined by something other than Anne. Since only Anne's mental set bears on the choice she will make, no other factors play a causal role in bringing about what she will choose. On this account of luck, then, it is false to say that her choice is lucky. So how the putative links are developed between indeterminism, luck, and freedom (or its absence) matters for evaluating whether or not luck is a threat to libertarians.

4. The Rollback Argument

4.1 The Initial Argument

Peter van Inwagen (2000) uses what has become known as the 'Rollback Argument' to develop the luck problem for libertarians.[4] Van Inwagen asks us to imagine that Anne's universe is 'rewound' by God to right before she make her choice. Suppose God then

4. Van Inwagen presents the Rollback Argument as a way of developing what he terms the *Mind* argument. As van Inwagen explicates it, the *Mind* argument develops the idea that libertarian choices are not free because they are "mere matters of chance" and the Rollback Argument is a way of developing this intuitive idea (van Inwagen 2000). But as Franklin (2012) points out, van Inwagen earlier (1983) presented the *Mind* argument as having three separate instances; van Inwagen does not say which the Rollback Argument is supposed to develop or whether it is somehow supposed to develop the overall idea of all three. Franklin then wants to separate luck arguments and the *Mind* argument; treating the Rollback Argument as an instance of the first. I here do the same, reading van Inwagen's appeal to "mere matters of chance" as developing the concern regarding luck.

allows things to play out again. And then again, and again. Suppose that God rewinds the universe and causes it to replay 726 times. In about half of the replays, Anne chooses to intervene while in the other half she chooses punctual attendance at the meeting. After we observe the replays, van Inwagen comments:

> …we shall be faced with the inescapable impression that what happens in the seven-hundred-and-twenty-seventh replay will be due simply to chance…[W]hat other conclusion can we accept about the seven-hundred-and-twenty-seventh replay (which is about to commence) than this: each of the two possible outcomes of this replay has an objective, 'ground-floor' probability of 0.5—and there's nothing more to be said? And this, surely, means that, in the strictest sense imaginable, the outcome of the replay will be a matter of chance. (2000, 15)

He continues,

> If [Anne] was faced with [two options], and it was a mere matter of chance which of these things she did, how can we say that—and this is essential to the act's being free—she was *able* to [stop and help] and *able* to [go to the meeting]? How could anyone be able to determine the outcome of a process whose outcome is a matter of objective, ground-floor chance? (2000, 15–16 italics original)

The implied answer is clear: no one is able to determine the outcomes of such a process. I thus interpret van Inwagen as here providing an argument schema similar to the 'luck' schema presented by Franklin above. This argument runs

(i) If an action is undetermined, then it is a *mere matter of chance*.
(ii) If an action is a *mere matter of chance*, then it is not free.

Thus, I take van Inwagen to be offering an account of luck in terms of mere chance and then claiming that rollback scenarios show that rolled-back choices are matters of mere chance and thus not free. Importantly, while the claim that Anne's choice is a mere matter of chance gets significant intuitive support from consideration of the rollback scenarios, the key point is that her choice is a mere matter of chance even on the very first scenario. That is, van Inwagen presents the rollbacks as a way of intuitively showing that when Anne is faced with the choice in the *actual* world, it is a mere matter of chance that she chooses to stop and help the assault victim.

4.2 Franklin's responses

Christopher Franklin has recently responded to the Rollback Argument on behalf of libertarians. One strand of reply invokes the following account of abilities:

> An agent *S* has the ability to *Φ* at *t* in *W* only if there is a set of possible worlds *w*, that is such that, all the worlds in this set have the same laws of nature as *W*, *S*'s intrinsic properties are sufficiently similar to her intrinsic properties in *W*, and *S Φ*-s. (Franklin 2011, 218)

This account of abilities is plausible and I additionally grant it to Franklin for the sake of argument. The idea is that if we understand what it is for an agent to have an ability, we will see that Anne has both the ability to stop and help and also the ability to go to the meeting. Thus, Franklin's account of abilities allows him to answer van Inwagen's first question in the second block quote above: "how can we say that—and this is essential to the act's being free—she was *able* to [stop and help] and *able* to [go to the meeting]?" We can say that she is able to stop and help because there are many worlds with the same laws of nature as the actual world, Anne's intrinsic properties in those worlds are sufficiently similar to her properties in the actual world, and Anne stops to help. And we can say that Anne has the ability to go to the meeting for the same reasons.

With this strand of argument, then, Franklin presents an account of abilities that grounds the claim that Anne has both the ability to help and the ability to go to the meeting. He admits that it is still undetermined that Anne *exercises* her ability to stop and intervene in the assault (2011, 218). But he believes that "we are left with little to no reason for thinking that indeterminism introduces a kind of luck or chance that is incompatible with an agent…being free and morally responsible" (2011, 218–219).

Franklin has met van Inwagen's initial challenge. Recall that part of van Inwagen's challenge was to answer the question "how can we say that—and this is essential to the act's being free—she was *able* to [stop and help] and *able* to [go to the meeting]?" We should agree that Franklin's account of abilities shows that Anne is able to do both. Adding further support for this claim, Franklin argues (2012) that rollback scenarios are just a way of demonstrating indeterminism, and thus don't present a significant challenge to libertarianism.[5] His idea is that the rollback scenarios simply show what it would

5. Franklin also notes that van Inwagen's initial description of the rollback thought experiment is metaphysically impossible. That is because, as van Inwagen describes, God continually rolls back time in the same possible world, but we are asked to imagine that sometimes Anne makes different decisions in the future of that world. But Anne cannot make different decisions regarding the same choice at the same time period in the same world, as "a possible world has all its components essentially: a possible

mean for Anne's choice not to be determined and thus only "describe libertarianism in a rather colorful way. But one cannot raise the cost of libertarianism by simply describing it" (2012, 409). However, it is not clear that Franklin has fully vindicated libertarianism against the Rollback Argument.

4.3 Schlosser's reply

While Franklin is able to demonstrate that Anne possesses the dual abilities in question, the success of his ultimate reply to the luck problem has been questioned by recent work. Let us grant that Anne has both the ability to help and the ability not to help. Does that show that Anne's choice—the choice she makes in the actual world—is truly free? The worry about freedom is a worry about control. If Anne stops and helps, is she in sufficient control of her choice? To answer this question in the affirmative it may not be enough to show that whichever way Anne chooses she will have chosen for reasons and that she was not coerced or compelled—as alluded by Kane when discussing plural voluntary control. Markus Schlosser has recently argued that answering 'yes' to the question requires that Anne have the power to choose one alternative rather than another. On Schlosser's view, the real challenge the Rollback Argument presents to libertarianism is to give an account of how Anne has control sufficient to exercise her dual ability in one way rather than another. While she can either stop and help or go to the meeting, she cannot "exercise either one of the two abilities *such that* she can *select* which alternative to pursue" (Schlosser 2014, 381 emphasis original).

Seeing this might appear to show that, contra Franklin's assertions, the Rollback Argument does raise the cost of libertarianism. And it does that even though there is a sense in which the thought experiment simply shows a vivid demonstration of what indeterminacy between a person's mental set and her choice involves. If her choice is not determined then sometimes she could, and would, choose the other way rather than the way she actually does choose. That much is, and should not be, in dispute. The real threat of the Rollback thought experiment to libertarianism, however, concerns what it implies about *Anne's original choice*. It shows that while she has the ability to choose either way

world could not have been different" (Franklin 2012, 407). However, this is not a serious barrier to consideration of the rollback idea, because as Franklin notes, we can imagine instead God rolling back time and letting Anne decide again. Any decision that is different will thus take place in a *different possible world*. Technically, then, rollback scenarios rollback time to a possible world that has more than one world as direct 'descendant.' As time rolls forward the world 'branches' into at least two sets of worlds, one set where Anne stops and helps and the other where she continues on to the meeting.

she does not have sufficient control over the original choice itself. If it is simply a matter of objective chance that she chooses to stop and help rather than continue on to the meeting, she appears as much in control of her decision as she would have been if the way she chose were simply determined by a coin flip. Since Anne would not be in control when the coin-flip selects the option of stopping and helping, why think that she is in control when her act is the result only of her own mental set? Franklin has not given us an answer to this question, which I think is the real question forced on libertarians by the Rollback Argument.[6]

5. A Way Forward

5.1 Suggestive Return to Kane and van Inwagen

Kane develops the notion of plural voluntary control in conjunction with his account of *self-forming actions* (SFAs). I am worried that Kane's focus on the importance of *self-forming actions* (SFAs) may have led others astray in thinking about libertarian models of control. According to Kane,

> SFAs occur at those difficult times of life when we are torn between competing visions of what we should do or become. Perhaps we are torn between doing the moral thing or acting from ambition, or between powerful present desires and long-term goals, or we are faced with difficult tasks for which we have aversions. In all such cases, we are faced with competing motivations and have to make an effort to

6. John Fischer has recently responded to the Rollback Argument in a way that might appear to address this worry (2012; forthcoming). He suggests we imagine that someone is morally responsible for her choice to raise her hand. Hold fixed the supposition that she is morally responsible for doing so, then imagine that we add to the description of her case a machine that 50% of the time will do nothing, but 50% of the time will stimulate her brain to cause her to refrain from choosing to raise her hand. Because the operation of the machine is random, we can run the rollback scenarios and see that 50% of the time the person raises her hand, while 50% of the time she does not. But, Fischer urges, if we supposed the person was morally responsible for raising her hand in the first place, we should still consider her morally responsible once we add the machine even though the machine makes it indeterminate that she will raise her hand.

I worry that Fischer's strategy does not fully address the argument because when the machine operates it *preempts* the person's choice: 50% of the time she chooses to raise her hand while 50% of the time the machine directly stimulates her brain to prevent her from choosing. So while the objective probabilities are the same as in Anne's case, there is a crucial difference from van Inwagen's rollback scenarios. In the rollback scenarios the worry about control emerges because it is clearly the agent, herself, who chooses differently.

> overcome temptation to do something else we also strongly want. (2007, 26)

Kane's idea is that libertarians do not need to require indeterminacy between a person's mental set and her choices at all times. They only need sufficient indeterminacy in the right place to ground an agent's ultimate responsibility for what she is like. Essentially, Kane's thought is that if an agent can be found ultimately responsible for her mental set then she will also be responsible for any choices that flow from that set.

SFAs are what Kane uses to ground that ultimate responsibility. As he notes, "In SFAs, the agent's will is divided and the agent has strong reasons or motives for making either choice" (2007, 29). In these cases, when our motives and reasons are balanced, "we *make* one set of competing reasons or motives prevail over the others then and there *by deciding*" (2007, 26–27). By making one set of reasons prevail over the others, we make ourselves; we impact the makeup of our mental set. And then it is by virtue of our responsibility for our mental set that we are responsible for all other choices. But SFAs only concern cases where our motives are balanced. For Kane, then, it is key that in SFAs the indeterminacy of a person's actions is reflective of a balancing of motives.

Consider, now, van Inwagen's initial presentation of the idea of rolling back, or replaying, an agent's choice. When presenting the Rollback Argument he notes that

> We may, for example, observe that, after a fairly large number of replays, Alice lies in thirty percent of the replays and tells the truth in seventy percent of them—and that the figures 'thirty percent' and 'seventy percent' become more and more accurate as the number of replays increases. (2000, 14)

But he then goes on to imagine the "simplest case": the case where each choice occurs 50% of the time. Why? Well, one thought is that if the simplest case is sufficient to make the point there is no need to consider a more complex case. But another is that given the importance of Kane's SFAs in the literature, the simplest case is the most important one. Whatever the reasons, my concern is that a focus by libertarians and their critics on the simple case has made libertarians less able to respond to worries about luck. We don't yet have an account of how an agent can exhibit significant control if, in rollback scenarios, she acts differently 50% of the time. My strategy is to suggest that libertarians need to get further inside the heads of the relevant agents to respond to the worry about luck presented by the Rollback Argument. We need to say more about the mental sets of the

agents in question in order to more fully understand how they can be in control over their choices.

5.2 Anne and Jan

Let's return to Anne; I want to now reply to Schlosser's concern that Anne lacks sufficient control. I admit that Anne, herself, is not able to select exactly which alternative to pursue in the actual world. (Acknowledging this is just what Franklin has referred to as 'describing libertarianism.') But I think there is room for libertarians to argue that Anne possesses as much freedom and control over what she does as is possible, so long as the evaluative elements of her mental set are equally inclining her toward either choice. This last fact is crucially important.

Consider, by contrast, Jan instead of Anne. Jan is also a successful businesswoman on the way to an important meeting. Like Anne, Jan happens upon someone being assaulted. She must decide whether to intervene and help the victim of the assault or continue on to her meeting. And like Anne, Jan's choice is to stop and intervene. Finally, like Anne, the link between Jan's mental set and her choice is indeterministic. Given her mental set at the time of her choice, it is not ensured that she will choose to stop and help.

Just like Anne, God 'rolls back' Jan's choice so we can see how she would choose in alternative scenarios. And again like Anne, we discover that Jan's choices are roughly split between the two alternatives as the scenarios unfold. The crucial difference, however, is that the evaluative elements of Jan's mental set vastly favor stopping and helping the assault victim over going to the meeting. But Jan is weak-willed, so her evaluation is not reflected in her pattern of choice, which obeys the simple case's 50%-50% split.

What do I mean by the evaluative elements of Jan's mental set? Well, suppose that Jan judges that it is best for her to stop and help the assault victim. Perhaps it *also* turns out that she has resolved in the past to help people in need even if it means forgoing important benefits to herself.[7] (Maybe Jan worries that she is too quick to favor her own interests over the needs of others when the temptation arises. Her resolution reflects her commitment to change.) But in spite of these facts, Jan's actions in the rollback scenarios often also reflect her desire for the potential promotion she could secure via attendance

7. In a recent paper (2012), Joshua May and Richard Holton argue that the ordinary concept of weakness of will is a prototype, or cluster, concept that involves both acting contrary to best judgment and also too quickly revising a previously-made resolution. I try here here to include both elements in Jan's mental set. For more on resolutions, see (Holton 1999).

at the meeting. But this desire for progress up the corporate ladder is one that she repudiates, has resolved not to act on, and works actively to extinguish—suppose Jan is not proud of her attraction to status and the increased salary isn't worth the extra responsibility. Regardless of the particular explanation(s), we find Jan's pattern of action displayed in the rollback scenarios to not be well predicted by the evaluative elements of her mental set.[8] Jan, I submit, is thus significantly out of control compared to Anne.

Anne's comparative control is explained by the fact that Anne is genuinely torn about what to do in the situation. She takes herself to have about equal reason to go to the meeting as to stop and help the victim of the assault. Perhaps she has also resolved to try to balance her career ambitions with her desire to help others from time to time. Thus, though Anne only helps the victim 50% of the time in the rollback scenarios, she does not display weakness of will in doing so. I take it that means Anne also does not display weakness of will in the *actual* scenario. When Anne decides to stop and help the victim, she does so for reasons she has and she endorses. They are not reasons she finds to be particularly overriding, of course. Anne would not be shocked at herself if in a similar future scenario she did not stop and help. Additionally, Anne is not coerced or compelled to stop and help. Anne has all the control over her act an agent can be expected to have.

In contrast, Jan would be horrified to discover about herself that she only helps the assault victim 50% of the time in rollback scenarios. She wholeheartedly judges that her minor status ambitions should take a backseat to helping others in sufficient need. Further, she has resolved to never fail to help others even if there is an enticing career prospect in play. Given these facts, knowing that she only stopped to help 50% of the time would shock her. (Or, at least, it would shock her if she also thinks of herself as a mostly continent person.) And these facts about Jan should bother us, too. We see exhibited in Jan a defect of agency—of agential *control*—which is not present in Anne.

Compare yet another agent, Stan, to both Anne and Jan. Stan is strongly committed to helping the assault victim in the actual scenario, just like Jan. He judges that it is clearly best for him to help and he has also resolved to always help in scenarios like this. When we rollback Stan's universe, however, we find that Stan chooses to help the assault victim 978 times, while he hurries on to the meeting only 22 times. Stan's choice is still not determined by the interaction of his prior mental set and the world. After all, he can

8. If Jan's resolution coupled with her judgments about what is best for her do not pick out the relevant elements of Jan's mental set that 'truly stand for her,' I invite the reader to substitute her own favored notions instead. That is, imagine whatever needs to be true for Jan to be strongly committed to acting one way, but at the same time, it is consistent that Jan often acts in the contrary manner.

and sometimes does choose to attend the meting over helping the assault victim. But since Stan chooses to help the assault victim 97% of the time, libertarians can hold—correctly—that Stan exercises significantly more control over his choice than Jan does.

Suppose that the right thing to do is to stop and help the victim of the assault. If so, then not only does Stan exercise more control over his choice to stop and help, his decision to stop and help is more praiseworthy than Anne's. The fact that he chooses to stop and help 97% of the time demonstrates both the strength of his moral concern for people unjustly victimized by others and his continence in choosing in such contexts. In contrast, neither Anne's nor Jan's choice-pattern reflects particularly well upon her. The important thing, however, is that it is for very different reasons. Anne's choice pattern reflects that she is not particularly concerned for those in need (at least, when there are payoffs for her), while Jan's reflects that she has significant weakness of will (at least, in this context).

Libertarians require an indeterministic link between an agent's mental set and her choices (at least, at certain key points in the life of an agent). What I am suggesting is that while the causal relation between an agent's mental set and her choices must be indeterministic, there is no reason that the indeterministic relation is always one where rollback scenarios show the agent choosing either option 50% of the time. What is more important, I urge, is the degree of fit between the outputs of the evaluative elements of an agent's mental set and her overall patterns of action. When the agent really is torn between two choices, her patterns of action in rollback scenarios should reflect that. If she is not on the fence, she should choose one option significantly more often when time is rolled back. My thought, then, is that libertarians might *embrace* rollback scenarios as potentially revealing important facts about an agent that impair agential control. Sometimes what rollback demonstrates weakens control, namely when the percentage of times the agent acts in a particular way does not reflect the degree to which she is committed to that option. However, sometimes it does not. To adequately respond to the Rollback Argument, then, libertarians need to talk more about agent's commitments to the various courses of action they consider.

5.3 Degrees of Control

I have argued that rollback scenarios need not harm libertarians. If we are more specific about the mental sets of the agents we consider, libertarians can use rollback scenarios to help explicate an agent's degrees of control and the praise or blameworthiness of her choices. My idea here relates libertarians to a somewhat unlikely ally. John

Martin Fischer and Mark Ravizza (1998) develop a similar account for *compatibilists*. On their model, an agent's control is explicated—roughly speaking—by appealing to counterfactuals about how the agential mechanism of a person's choice would perform in relevantly similar possible worlds. Thus, rollback scenarios are the libertarian counterpart to Fischer and Ravizza's compatibilist idea. Rollback scenarios are counterfactuals about how an agent would act in the *same world*, rather than relevantly-similar worlds.[9]

So rollback scenarios can do two helpful things for libertarians. First, they can show the degree of control an agent exercises over her action. To determine this, we ask in the rollback scenarios whether there is an appropriate mesh between the percentage of the time the agent chooses one option and the strength of her evaluative commitment to it. Thus, Anne and Stan exhibit more control than Jan, because their choices in the rollback scenarios comport with their evaluation of the desirability of the options. Jan's choices, by contrast, do not fully reflect where she stands on the issue confronting her.

Second, rollback scenarios can help to show the degree of praise and blameworthiness an agent bears for her action. Consider, then, the relative degree of praiseworthiness each agent—Anne, Jan, and Stan—bears for the act of deciding to stop and intervene in the assault. (Remember, in the actual world, all three stop and help.) Sometimes Anne helps, sometimes Jan helps, and sometimes Stan helps. However, in the rollback scenarios Anne and Jan each helps about 50% of the time, while Stan helps 97% of the time. Thus, Stan's choice to help is more praiseworthy as it is more reflective of both a substantial resolution to aid when needed and his judgment that helping is the best thing for him to do in the situation. Assuming that stopping to help is the right thing to do, Anne's choice to help is not nearly as praiseworthy, because she is only moderately in favor of helping in such scenarios. Similarly, Jan's choice is not significantly praiseworthy, but for a different reason: her lack of control. We have reason to withhold praise either when a person is not strongly committed to what is morally right or when she lacks the ability to exhibit her commitment to what is right in action. Both are moral defects, though of very different kinds.[10]

9. Potentially, this could provide a small advantage for libertarian accounts of free agency, as libertarians do not then need to invoke or define which worlds are the relevant ones.

10. We might also have reasons to praise a weak-willed person: i.e. reasons to praise someone who is not praiseworthy. Perhaps, for example, praising the weak-willed person will encourage her to be more continent. Such considerations would take us far afield of the current topic, which is when people really deserve praise and blame for their choices such that they are morally responsible for them. For more, see (Cogley 2013).

At this point, an objector might urge that there is nothing inherently *libertarian* about my response to the Rollback Argument. After all, consider an agent who is *totally* in favor of one of the options she confronts. Suppose Anne happens on a person in need and at the same time has a very strange thought that she could instead go get ice cream. Puzzled, she rejects this thought: she's all in for helping in this scenario. Thus, perfect continence and control for Anne in this case would be exhibited if in rollback scenarios she always chooses just that option; 100% of the time she chooses to help the person in need. But this would just be for Anne's action to be determined by her mental set: anathema to a libertarian.

Certainly, on the account I am developing, a libertarian must require an indeterministic link between a person's mental set and her choice. So the libertarian must balk at attributing full control when rollbacks show a person doing the same action 100% of the time. I do not have space to defend the claim that actually having leeway between options enhances a person's control over her choices. That is a fundamental libertarian commitment which I am simply assuming here. What I've tried to do is show that a failure to make the same choice 100% of the time in rollback scenarios is very much compatible with someone exhibiting significant control over what she does. I've thus provided a defense for libertarians against the rollback version of the luck problem, which is a problem about indeterministic agents having diminished control over their choices. Whether indeterministic agents have enhanced control over their choices compared to fully determined agents is another topic.

6. Conclusion

I've here sketched an account of how libertarians can respond to worries about luck presented by rollback scenarios. My thought is that discovering that in rollback scenarios someone would act differently than she actually does 50% of the time need not make us think the agent lacks substantial control over what she does. Libertarians can, and should, insist that control is fundamentally about the core evaluative elements of a person's mental set being translated into choice. If we are clear about the nature of a person's mental set and that she really is torn between the options, then finding out that she chooses differently 50% of the time just is to see her continence demonstrated via a divine mechanism.

Even if an agent's choice is undetermined, the choice may still be free and under her control if affected by the strength of the agent's commitment to the various courses of action. Counterfactual rollback scenarios that show how the agent could have acted if

the world were rewound to the exact state and time of her choice might then be a way of exhibiting the agents commitment and continence. The fact that an agent would have acted differently in such scenarios is thus consistent with her having control over her action sufficient for her to be free and morally responsible for it.

References

Clarke, Randolph. 2003. *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Cogley, Zac. 2013. "Basic Desert of Reactive Emotions." *Philosophical Explorations* 16 (2): 165–77.

Fischer, John Martin. forthcoming. "Toward a Solution to the Luck Problem." In *Libertarian Free Will: Contemporary Debates*, edited by Robert Kane. New York: Oxford University Press.

Fischer, John Martin. 2012. "Indeterminism and Control: An Approach to the Problem of Luck." In *Deep Control*. New York: Oxford University Press.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, Cambridge University Press.

Franklin, Christopher Evan. 2011. "Farewell to the Luck (and Mind) Argument." *Philosophical Studies* 156 (2): 199–230. doi:10.1007/s11098-010-9583-3.

Franklin, Christopher Evan. 2012. "The Assimilation Argument and the Rollback Argument." *Pacific Philosophical Quarterly* 93 (3): 395–416. doi:10.1111/j.1468-0114.2012.01432.x.

Holton, Richard. 1999. "Intention and Weakness of Will." *The Journal of Philosophy* 96 (5): 241. doi:10.2307/2564667.

Kane, Robert. 1996. *The Significance of Free Will*. New York: Oxford University Press.

Kane, Robert. 2007. "Libertarianism." In *Four Views on Free Will*. Oxford: Blackwell.

May, Joshua, and Richard Holton. 2012. "What in the World Is Weakness of Will?" *Philosophical Studies* 157 (3): 341–60. doi:10.1007/s11098-010-9651-8.

Schlosser, Markus E. 2014. "The Luck Argument against Event-Causal Libertarianism: It Is Here to Stay." *Philosophical Studies* 167 (2): 375–85. doi:10.1007/s11098-013-0102-1.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. New York: Oxford University Press.

Van Inwagen, Peter. 2000. "The Eighth Philosophical Perspectives Lecture: Free Will Remains a Mystery." *Noûs* 34 (October): 1–19.

Journal of Cognition and Neuroethics

Agential Settling Requires a Conscious Intention

Yishai Cohen
Syracuse University

Biography
Yishai Cohen is a PhD candidate in the philosophy department at Syracuse University. His research interests lie primarily at the intersection of agency and metaphysics.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Cohen, Yishai. 2015. "Agential Settling Requires a Conscious Intention." *Journal of Cognition and Neuroethics* 3 (1): 139–155.

Agential Settling Requires a Conscious Intention

Yishai Cohen

Abstract
Helen Steward holds that an agent's settling something does not require a conscious, full-fledged intention. Rather, sub-intentional acts can be instances of settling by the agent if that act is subordinated to the agent's personal-level conscious systems. I argue that this position is mistaken, and that agential settling does in fact require a conscious intention. I argue for this claim by offering a case which on Steward's position has counterintuitive implications. I consider a variety of ways in which Steward might respond, and show how each response incurs serious dialectical burdens. I then propose my preferred view of agential settling which does not share the aforementioned counterintuitive claims.

Keywords
Helen Steward, determinism, compatibilism, incompatibilism, libertarianism, settling, agency, intention

Introduction

In *A Metaphysics for Freedom*, Helen Steward argues that determinism is false on the basis of her notion of settling, a notion which is the central theme of her book. The argument may be summarized as follows:

1. If determinism is true, no one settles anything.
2. Humans and nonhuman animals settle things.
3. Therefore, determinism is false.[1]

1. Steward's actual argument against determinism is more complicated than what is being presented here, though these further complications do not concern what I wish to argue for in this paper. Regardless, Steward's (2012, 12) actual argument against determinism is as follows:

    1. If universal determinism is true, the future is not open.
    2. If there are self-moving animals, the future is open.
    3. There are self-moving animals.
    4. Therefore, universal determinism is not true.

    Self-moving animals possess the capacity to move their body whereby their "contribution *does* amount to something over and above the contribution of the process inside them which eventuate in the resulting bodily movements" (2012, 16–17).

Steward (2012, 39–42) distinguishes between a weak and strong account of settling, whereby the strong account is employed in premises (1) and (2). According to the weak account of settling, there is not necessarily any privileged time at which some event *e* is settled.[2] Rather, *e* can be settled by multiple events that occur at different times, so long as these events are part of the causal chain that lead up to *e*'s occurrence. To illustrate, on the weak account of settling, if determinism is true, then the state of the world in the remote past in conjunction with the laws of nature settle that my arm rises at *t*. However, my decision in a deterministic world to raise my arm *also* settles that my arm rises at *t* since my decision is part of the causal chain leading to my arm's rising at *t*. In other words, an overdetermination of settling by events that occur at distinct times is possible on the weak account of settling. By contrast, on the strong account of settling, an overdetermination of settling by events that occur at distinct times is impossible.

According to the strong account of settling, if at time t_1 it is nomologically possible that *e* occur at t_4, and at t_1 it is nomologically possible that *e* not occur at t_4, then at t_1 it is not settled whether *e* will occur t_4. Let's further suppose that at t_2 it is nomologically possible that *e* occur at t_4, but that at t_2 it is nomologically impossible that *e* not occur at t_4. In that case, whether *e* occurs at t_4 is settled in the strong sense by some event at t_2. Moreover, given that an overdetermination of settling by events that occur at distinct times is impossible on the strong account of settling, since some event at t_2 settles that *e* will occur at t_4, no event that occurs at t_3 can settle that *e* occurs at t_4, *including* events at t_3 that cause the occurrence of *e* at t_4.[3]

Given that Steward is employing the strong account of settling in her argument against determinism, premise (1) is undeniably true, and the crucial and controversial premise is (2). My disagreement with Steward does not concern the truth of (2). Rather, my aim is to express an in-house disagreement between myself and Steward concerning the finer details of the strong account of settling (all subsequent discussion of settling solely concerns the strong account of settling). More specifically, Steward claims that settling does not require a conscious, full-fledged intention. Rather, sub-intentional acts can be instances of settling by the agent if that act is subordinated to the agent's personal-level conscious systems. It is this position I wish to dispute.

2. Strictly speaking, Steward holds that that which is settled is not some event *e*, but rather a question of whether-*p*, whereby *p* may refer to the proposition that event *e* occurs (at some time). I will continue to speak of events rather than questions being settled merely for brevity's sake, as this will make no difference to the discussion below.

3. For further discussion of these two accounts of settling, see Clancy (2013).

This paper is divided into four parts. In section 1, I argue that, contra Steward, agential settling does in fact require a conscious intention. I argue for this claim by offering a case which on Steward's position has counterintuitive implications. In section 2, I consider a number of ways in which Steward might reply to my case. I attempt to show that these replies do not succeed insofar as each reply incurs serious dialectical burdens. In section 3, I propose my preferred view of settling. Finally, in section 4 I argue that my view does an equally good job of supporting Steward's argument against determinism, and, moreover, that my view can offer a more satisfying answer to the luck argument against libertarianism.

1. A Problem for the Subordination Thesis

As previously noted, Steward maintains that settling does not require an antecedent (or simultaneous) conscious intention (2012, 47).[4] To illustrate, consider an agent *S*'s sub-intentional act such as *S*'s head slightly turning or *S*'s foot jiggling which are not produced by means of *S*'s conscious intentions. Steward (2012, 50–52) maintains such a sub-intentional act by *S* is nevertheless an instance of settling by *S* if *S*'s sub-intentional act satisfies a certain condition which is captured in the following thesis:

> ***The Subordination Thesis (ST)*** Agent *S*'s sub-intentional act *A* is settled by *S* if *X* is causally responsible for the occurrence of *A*, and *X* is subordinated to *S*'s personal-level conscious systems insofar as *S* can consciously alter or prevent altogether the occurrence of *A*.[5]

My argument against *ST* appeals to what I'll call a *Reverse Frankfurt* case in light of the fact that my case reverses some of the structural features of Frankfurt's (1969) case against the principle of alternative possibilities.[6] Before I present this case, however, I must first present Jennifer Hornsby's (1980) distinction between transitive and intransitive verbs—a distinction which Steward herself accepts and employs. Roughly,

4. Steward does not say explicitly whether she understands a decision to be the formation of an intention, or simply the intention itself. Regardless, in light of affirming the subordination thesis, Steward clearly thinks that agential settling requires neither an intention nor the formation of an intention.

5. In Steward (2009), although the notion of settling is not employed, Steward similarly defends the position that sub-intentional actions are things we do—they are actions.

6. In Frankfurt's (1969) case, an agent allegedly cannot do otherwise given the presence of a preemptive intervener, even though the preemptive intervener stands idly by and never in fact intervenes. By contrast, in my *Reverse Frankfurt* case, an agent *can* do otherwise, even though an intervener *is* intervening.

a transitive verb refers to someone's *doing* something, and an intransitive verb refers to something *happening*. Here are two examples: the chair moved$_I$ (something happened to the chair) because I moved$_T$ the chair (I did something). My foot jiggled$_I$ (something happened to my foot) because I jiggled$_T$ it (I did something). With this distinction in hand, I now present my *Reverse Frankfurt* case:

> ***Reverse Frankfurt*** Jones is concentrating intensely on an exam she is currently taking, and has no interest in focusing her attention elsewhere. Unbeknownst to Jones, Black has placed a computer chip in Jones' brain which in turn might cause Jones' foot to jiggle$_I$ if Black presses a button. However, this chip is causally inert when it conflicts with Jones' conscious intentions: if Jones forms an intention to refrain from jiggling$_T$ her foot, the jiggling$_I$ will not occur even if Black presses the button. Hence, the chip is subordinated to Jones' personal-level conscious systems insofar as Jones can consciously alter or prevent altogether the jiggling$_I$.
>
> Given the fairly uncontroversial assumption that simultaneous causation is metaphysically possible (Taylor 1966; Brand 1980; Huemer and Kovitz 2003), the following is true:[7] Black's pressing of the button at time *t* deterministically causes the chip in Jones' brain to activate at *t*. Additionally, if at *t* the chip in Jones' brain activates and Jones has not formed an intention to refrain from jiggling$_T$ her foot, then the chip deterministically causes Jones' foot to jiggle$_I$ at *t*. So, at time *t* the following occurs:
>
> - Black presses the button.
> - The chip in Jones' brain activates.
> - Jones does not form (and has not formed) an intention to refrain from jiggling$_T$ her foot.
> - Jones' foot jiggles$_I$.

7. Steward would surely not want her account of settling to be committed to the metaphysical impossibility of simultaneous causation. At any rate, such a commitment would certainly seem to be a cost to her view.

ST renders the verdict that Jones' jiggling$_{I}$ is settled by Jones. Moreover, *ST* also seems to suggest that Black's pressing of the button does not settle that Jones' foot jiggles$_{I}$ precisely because the jiggling$_{I}$ is settled by *Jones*. For, it would appear that only one agent can settle some event if we wish to maintain that an agent settles some event if and only if it is up to the agent whether the event in question occur. At any rate, irrespective of what Steward might say about Black, her commitment to *ST* commits her to the claim that Jones' jiggling$_{I}$ is (at least) settled by Jones. This verdict is counterintuitive. Jones is concentrating intensely on an exam and is paying no attention to how her body might move in some trivial manner. As a result, if Jones' jiggling$_{I}$ is settled by an agent, it is at best settled by Black rather than Jones. In light of *ST*'s counterintuitive implications, I will now consider two ways in which Steward might attempt to modify *ST* in order to render the intuitively correct verdict in *Reverse Frankfurt* that Jones' jiggling$_{I}$ is not settled by Jones.

2. Revising the Subordination Thesis

In order to render the intuitively correct verdict in *Reverse Frankfurt* while *also* maintaining that in ordinary circumstances we often settle how our body moves in the absence of a conscious decision or intention, Steward may wish to modify *ST* in the following manner:

> ***The Subordination Thesis 2* (*ST2*)** Agent *S*'s sub-intentional act *A* is settled by *S* if *X* is causally responsible for the occurrence of agent *S*'s sub-intentional act *A*, and *X* is subordinated to *S*'s personal-level conscious systems insofar as *S* can consciously alter or prevent altogether the occurrence of *A*, *and X does not involve in any direct way the intentions of other agents*.

ST2 is meant to be understood in such a way that Jones' jiggling$_{I}$ is not settled by Jones because Black's intentions are involved in some direct way with Jones' jiggling$_{I}$. The notion of 'not involving in any direct way' is, among other things, undoubtedly vague. But never mind that. There are two more urgent problems with *ST2*.

First, a worry arises that affirming *ST2* would thereby undermine the manipulation argument which is one of the more powerful arguments for incompatibilism. In a nutshell, that argument claims that if *S*'s action is manipulated by other agents, then *S* is not morally responsible[8] for that action. Moreover, there is no relevant difference

8. The kind of responsibility at issue here is basic desert responsibility. To be responsible in the basic desert

between a case of manipulation and an ordinary case in which one acts in a deterministic universe. So compatibilism is false (Pereboom 2014, §4). If Steward accepts *ST2*, however, this opens up a 'soft-line' reply to the manipulation argument, according to which there *is* a relevant difference between an instance of manipulation and an ordinary case in which one acts in a deterministic universe. The difference is that only in an instance of manipulation are one's actions causally determined by factors beyond one's control, *whereby such factors involve in some direct way the intentions of other agents* (Lycan 1997). While I don't find this response to the manipulation argument compelling, it appears that if Steward adopts *ST2*, then she cannot consistently object to this soft-line reply to the manipulation argument. For, both a proponent of *ST2* and a proponent of the above soft-line reply accept importantly similar claims. One accepts that the intentions of other agents can make a difference with respect to whether *S* settles something. The other accepts that the intentions of other agents can make a difference with respect to whether *S* is morally responsible for what she has done.

The second problem with *ST2* is that, like the above soft-line reply, it seems *ad hoc*. Suppose that Black is replaced with a spontaneously emergent robotic machine that was not produced by an intelligent designer, and that the robotic machine causes Jones' jiggling$_{I}$ (Pereboom 2001, 115; 2014, 79). Alternatively, suppose that a spontaneously generated electromagnetic field directly causes the jiggling$_{I}$ (Mele 1995, 168–169; 2006, 141). Moreover, suppose that both the robotic machine and the electromagnetic field are subordinated to Jones' personal-level conscious systems. Surely these cases can't make the difference with respect to whether Jones' jiggling$_{I}$ is settled by Jones. *ST2* is thus untenable.

Steward might attempt to modify *ST* in a different way in order to escape the problems with *ST2*. Accordingly, such an amendment to *ST* must not invoke the notion of agency or action. For, it is plausible that such notions involve in a direct way the intentions of other agents. In that case, irrespective of the finer details of such an amendment, that amendment will entail the following:

> ***The Subordination Thesis 3* (*ST3*)** Agent *S*'s sub-intentional act *A* is settled by *S* if *X* is causally responsible for the occurrence of agent *S*'s sub-intentional act *A*, and *X* is subordinated to *S*'s personal-level conscious systems insofar as *S* can consciously alter or prevent

sense for performing an action is to deserve blame or credit just because one understands the moral status of the action one has performed, and not because of consequentialist or contractualist considerations (Scanlon 2013; Pereboom 2014).

> altogether the occurrence of *A*, *and X satisfies some further condition c, such that c does not invoke the notion of agency or action.*

ST3 seems to escape the charge of being *ad hoc* which I claimed plagues *ST2*. However, in the context of Steward's aims, *ST3* has its own serious problem, as I will now explain. Steward argues extensively against the causal theory of action (Davidson 1973; Frankfurt 1988; Bishop 1989; Velleman 2000), which Steward understands to be the thesis that "[f]or an agent to act is roughly…for the bodily movements that are intrinsic to the relevant action to be caused by certain of that agent's own mental states" (2012, 55). Steward endorses the two prevalent objections to the causal theory of action.

The first objection is that the *appropriate manner* in which an agent's mental states must cause one's bodily behavior in order to count as an action must be further specified. Davidson's (1973) classic illustration of this point involves a climber who is holding a rope to which another person is tied and who is endangering the climber. The climber wants to rid herself of this person, and could do so by loosening her grip of the rope. The climber's relevant beliefs and desires result in the climber's being extremely nervous, which in turn results in the climber *unintentionally* loosening her grip of the rope. In this example, the climber's loosening her grip of the rope was not an intentional action because her beliefs and desires did not cause the climber's bodily movements in the appropriate manner. This is the problem of deviant causal chains.

The second and closely related objection to the causal theory of action is that it in principle cannot provide necessary and sufficient conditions for when an *agent* does anything, i.e. where the agent—and not just her mental states—is responsible for the agent's relevant bodily movements. For, once the relevant mental states play their role in causing the agent's bodily movements, there's nothing left for the *agent* to do. This is sometimes called the disappearing agent objection (Hornsby 2004).

Now, Steward finds both objections to the causal theory of action persuasive, and, crucially, claims that neither objection can be answered so long as the proponent of the causal theory of action provides necessary and sufficient conditions for action which *do not invoke the notion of agency or action* (Steward 2012, 55–66). In other words, appealing only to the agent's mental states and their causal relation to the agent's bodily movements will always in principle leave out what we were after, viz. the *agent's* doing something. Steward invests a great deal in arguing against the causal theory of action precisely because it is a compatibilist-friendly one. So, it would be a great cost to Steward if she had to give up the two aforementioned objections to the causal theory of action. However, I claim that this is exactly what Steward must do if *ST3* is endorsed.

ST3 entails that there is some condition which is not met in *Reverse Frankfurt*, and yet *is* met in ordinary cases in which our sub-intentional acts are not directly caused by other agents, such that Jones' jiggling$_{I}$ is not settled by Jones, but an agent's sub-intentional act in an ordinary case *is* an instance of settling by the agent. The relevant event or property that satisfies this condition must surely involve a certain *way* in which one's mental states cause one's bodily movements. In other words, the way in which one's bodily movements are caused (by one's mental states) makes the critical difference between that which is, and is not, an instance of settling. But it seems to me that *ST3* opens up the door for the proponent of the causal theory of action to appeal to the exact same condition as making the critical difference between bodily movements that are and are not an action. Moreover, this condition is, *ex hypothesi*, one that is consistent with the causal theory of action precisely because this condition does not invoke any notions of agency or action. So I conclude that endorsing *ST3* would come at too great a cost for Steward given that her argument against the truth of determinism depends significantly upon refuting the causal theory of action. It seems, then, that there is good reason for Steward to give up *ST* and any variant thereof.

3. The Cartesian View of Settling

Since the proponent of Steward's argument against the truth of determinism should not endorse *ST* (or a variant of *ST*), I think the following position ought to be endorsed instead:

> ***Cartesian Settling* (*CS*)** Necessarily, an agent *S* settles some contingent event *e* only if *e* is preceded by (or is simultaneous with) a conscious intention by *S* to perform action *a*,[9] such that either (i) the occurrence of *a* at time *t* is identical to *e*, (ii) the occurrence of *a* at time *t* is identical to event *e** which necessitates *e*, or (iii) *a* deterministically causes *e*, and *S* believed that *a* might cause *e*.

Note that the above view is Cartesian only insofar as it emphasizes the role of the mind with respect to settling. This view is perfectly consistent with physicalism about the mind. For, nothing about *CS* requires that the conscious intention of the agent be a non-physical state. Moreover, this view is consistent with Steward's (2012, 16–17) view that agential

9. For simplicity's sake, I assume that an omission can be an action. If one disagrees, then *CS* could no doubt be appropriately tailored to accommodate such disagreement. I will further discuss the issue of intentional omissions below.

settling involves a kind of top-down causation that is over and above the causal processes that constitute the agent. To be clear, however, *CS* in no way *requires* top-down causation since *CS* is also consistent with a reductionist view of agency and settling (Franklin 2014).

In order to attain a better grasp of *CS*, let's consider the three kinds of ways in which an agent can settle something. Suppose Haley forms the intention to raise$_T$ her arm at *t* in order to get a taxi driver's attention, and Haley succeeds in raising$_T$ her arm and getting the taxi driver's attention. According to condition (i) of *CS*, Haley settles the occurrence of the following event: Haley's raising$_T$ her arm at *t*. Next, given the relationship between transitive and intransitive verbs, Haley's raising$_T$ her arm at *t* necessitates Haley's arm rising$_I$ at *t*. So according to condition (ii) of *CS*, Haley settles that Haley's arm rises$_I$ at *t*. Finally, consider the event of the taxi driver noticing Haley's signal. Haley's raising$_T$ her arm is, let's suppose, a deterministic cause of the taxi driver's noticing Haley's signal. Moreover, Haley believed that raising$_T$ her arm might cause the taxi driver to notice Haley's signal. So, according to condition (iii) of *CS*, it follows that Haley settled the occurrence of the taxi driver noticing Haley's signal.[10]

Now, in *Reverse Frankfurt*, Jones has no intention to jiggle$_T$ her foot, or to perform some action that entails that her foot jiggles$_I$, or to perform some action which Jones believes might cause her foot to jiggle$_I$. So *CS* renders the correct verdict that Jones does not settle her foot's jiggling$_I$. Additionally, since Black intentionally presses the button, the pressing of the button deterministically causes Jones' foot to jiggle$_I$ (in the absence of certain intentions by Jones), and Black believes that pressing the button might cause Jones' foot to jiggle$_I$, according to condition (iii) of *CS* it follows that Black settles that Jones' foot jiggles$_I$. So unlike *ST*, *CS* renders all of the intuitively correct verdicts in *Reverse Frankfurt*.

Before proceeding to the final section, I want to consider some important objections to *CS*. In order for an agent to settle something, certain factors outside of an agent's control intuitively need to obtain. For instance, suppose that Haley raised$_T$ her arm at time *t*, and there was a bomb nearby, such that it was nomologically possible for the bomb to explode at *t* (suppose that whether the bomb explodes at *t* depends upon certain genuinely indeterministic processes at the microphysical level). Call this *The Bomb Case*. Whether the bomb explodes at *t* is not up to Haley. Moreover, it is not up to Haley that if the bomb explodes, then Haley does not raise$_T$ her arm at *t*. We might be tempted to conclude that whether Haley raises$_T$ her arm at *t* is therefore not up to Haley. After

10. I have added a belief component to (iii) since I acknowledge that an agent's power to settle something partly depends upon an agent's beliefs (Shabo 2014).

all, this line of reasoning resembles a transfer of powerlessness principle employed in the consequence argument for the incompatibility of determinism and the ability to do otherwise (van Inwagen 1983). There is, however, a crucial difference between the above line of reasoning and the employment of a transfer of powerlessness principle. Haley's raising$_T$ her arm at *t* and not raising$_T$ her arm at *t* are each nomologically possible prior to *t*. Moreover, *if certain conditions beyond Haley's control which are nomologically possible obtain*, then it is up to Haley whether she raises$_T$ her arm at *t*. So, while Haley no doubt has a limited kind of control over whether she raises$_T$ her arm at *t*, we can nevertheless maintain that Haley settles that she raises$_T$ her arm at *t* in *The Bomb Case*. By contrast, if determinism is true then either it is nomologically impossible just prior to *t* (and at *t*) that Haley raise$_T$ her arm at *t*, or it is nomologically impossible just prior to *t* (and at *t*) that Haley not raise$_T$ her arm at *t*. So, *CS* is consistent with the view that no one settles anything (according to the strong account of settling) if determinism is true.[11] I now turn to another objection that arises in light of the remarks just made.

I have just claimed that *CS* is consistent with the following thesis:

> ***The Limited Settling Thesis*** Possibly, an agent *S* settles that *S* ϕ*-s* even when *S*'s ϕ-ing partly depends upon factors beyond *S*'s control.

A worry now arises that the proponent of *ST* can stand firm in asserting that *Jones* settles that Jones' foot jiggles$_I$ in *Reverse Frankfurt*. For, while Black is certainly a cause of Jones' foot jiggling$_I$, Black's intervention is still nomologically compossible with Jones' foot not jiggling$_I$. For, if Jones forms the intention not to jiggle$_T$ her foot, then Jones' foot will not jiggle$_I$, *irrespective* of whether Black intervenes. So, although whether Jones' foot jiggles$_I$ partly depends upon factors beyond Jones' control (given that Jones doesn't form an intention not to jiggle$_T$ her foot), it doesn't follow that the jiggling$_I$ is not settled by Jones.

In response, I contend that there is still a crucial difference between *Reverse Frankfurt* and *The Bomb Case*. In *Reverse Frankfurt*, Jones omits from jiggling$_T$ her foot. However, this omission is not *intentional* since Jones has no intention which is in any important way

11. I am intentionally side-stepping recent intricate issues concerning Joseph Campbell's (2007) 'no past' objection to the consequence argument, according to which the consequence argument does not demonstrate that an agent cannot do otherwise in a deterministic world since it is possible for an agent to perform an action at the first moment of time in a deterministic universe *without* being causally determined by factors beyond the agent's control to perform the relevant action. For a variety of responses to Campbell, see Brueckner (2008), Loss (2009; 2010), Bailey (2012), and Finch (2013). The present discussion can be appropriately tailored according to each reply.

relevant to Jones' foot not jiggling$_I$ (Clarke 2010). If, however, Jones *intentionally* refrained from jiggling$_T$ her foot, then Jones' foot would certainly not have jiggled$_I$, despite Black's intervention. Hence, while Jones' jiggling$_I$ in *Reverse Frankfurt* happens to partly depend upon factors beyond her control (viz. Black's intervention), and while Haley's raising$_T$ her arm in *The Bomb Case* likewise partly depends upon factors beyond her control (viz. the bomb not exploding), there is still a relevant difference between these two cases. Jones' jiggling$_I$ was not preceded by (or was simultaneous with) a relevant conscious intention by Jones. By contrast, Haley's arm rising$_I$ *was* preceded by (or was simultaneous with) a relevant conscious intention by Haley. It is *this* difference between *Reverse Frankfurt* and *The Bomb Case* that should lead us to conclude that Jones' jiggling$_I$ was not settled by Jones in *Reverse Frankfurt*, but Haley's arm rising$_I$ was settled by Haley in *The Bomb Case*. So, *The Limited Settling Thesis* should not lead us astray from the importance of intentions for agential settling. I now turn to the final section in which I show that *CS* does an equally good job of supporting Steward's argument against determinism, and, moreover, that *CS* can offer a more satisfying answer to the luck argument against libertarianism.

4. Cartesian Settling and Libertarianism

Steward emphasizes *ST* partly because she thinks *ST* poses a further obstacle to the compatibilist who adopts a causal theory of action. As we've seen, the causal theory of action is supposed to analyze actions in terms of some appropriate causal relation between the agent's mental states and her bodily movements. But sub-intentional acts are not causally produced by the agent's (conscious) mental states. So the causal theory of action must be false since it cannot accommodate sub-intentional acts (Steward 2012, 66–67).

If this was an intractable problem for the causal theory of action, then *ST* would have an advantage over *CS* insofar as only the proponent of *ST* can pose the aforementioned objection to the causal theory of action. But this problem *is* tractable. The proponent of the causal theory of action could revise her view in the following manner. An agent's sub-intentional act involves bodily movements that are causally produced by physical states within the agent. And these physical states are subordinated to the agent's personal-level conscious systems insofar as the agent's relevant mental states can modify or prevent altogether the agent's relevant bodily movements by nullifying the causal efficacy of the agent's relevant physical states.

This proposal analyzes the subordination of sub-intentional actions in terms of the causal efficacy of the agent's mental states. As a result, this proposal does not seem *ad hoc*, and moreover does not stray too far off from spirit of the causal theory of action which emphasizes the causal efficacy of an agent's mental states. So I conclude that Steward's objection to the causal theory of action with regards to sub-intentional acts fails, and thus that *ST* has no advantages over *CS* within the context of arguing for libertarianism. However, as I will now try to show, besides rendering the intuitively correct verdicts in *Reverse Frankfurt*, *CS* has a further advantage over *ST*.

There are a variety of luck arguments against libertarianism, the core of which is formulated by Franklin (2011, 201) as follows:

1. If an action is undetermined, then it is a matter of luck.

2. If an action is a matter of luck, then it is not free.

If (1) and (2) are true, undetermined actions cannot be free, and thus libertarianism is false. How might the libertarian respond? There is a trend in the literature to highlight the fact that a free action involves an *action* or that a free action is the *agent's* in order to undermine the luck argument as well as related arguments against libertarianism (Balaguer 2009; Franklin 2011; Griffith 2010). But there is reason to think that this response is inadequate, and that libertarians need something more, such as an account of how an agent can determine, select, or (to put it in Steward's terms) settle between the relevant multiple courses of action available to the agent (Schlosser 2014). More specifically, the libertarian arguably needs to adequately demarcate instances of agential settling from truly random outcomes (Shabo 2013).

In response to this challenge (and the luck argument) the *CS* proponent can say that, unlike truly random outcomes, agential settling *necessarily* involves a conscious intention, whereby doing something intentionally involves doing something *for a reason* (Davidson 1963; Goldman 1970; Mele 1992).[12] In other words, *CS* arguably implies that, unlike a truly random outcome, agential settling necessarily has a teleological explanation. I think this response has some merit (Lowe 2006; Goetz 2008). The trouble, however, for the *ST* proponent is that they cannot adopt this response. Sub-intentional acts do not require an intention on behalf of the agent. So they need not have a teleological explanation at all. But according to *ST* such sub-intentional acts can be instances of settling by the agent.

12. There are a variety of accounts—both causal and non-causal—of the relationship between an agent's action and her motivational reason for performing that action. I do not intend to take a stand on this issue here.

So, according to *ST*, agential settling does not necessarily have a teleological explanation. Hence, in comparison to the proponent of *ST*, the proponent of *CS* is better situated to respond to the luck argument (and related arguments) against libertarianism. More specifically, the *CS* proponent can point to a feature that is essential to agential settling which aids in demarcating agential settling from a truly random outcome. So I conclude that *CS* does an equally good job of supporting Steward's argument against determinism, and, moreover, that *CS* offers a more satisfying answer to the luck argument (and related arguments) against libertarianism. Steward thus has good reason to abandon *ST* in favor of *CS*.[13]

13. For helpful discussion and comments on a previous draft of this paper, I'm grateful to Kim Frost, Sean Clancy and the audience at the 2014 Free Will conference hosted by the Center for Cognition and Neuroethics.

References

Bailey, Andrew M. 2012. "Incompatibilism and the Past." *Philosophy and Phenomenological Research* 85 (2): 351–376.

Balaguer, Mark. 2009. *Free Will as an Open Scientific Problem*. Cambridge, MA: MIT Press.

Bishop, John. 1989. *Natural Agency*. Cambridge: Cambridge University Press.

Brand, Myles. 1980. "Simultaneous Causation." In *Time and Cause: Essays Presented to Richard Taylor*, ed. Peter van Inwagen, 137–153. Dordrecht: D. Reidel Publishing.

Brueckner, Anthony. 2008. "Retooling the Consequence Argument." *Analysis* 68 (1): 10–13.

Campbell, Joseph K. 2007. "Free Will and the Necessity of the Past." *Analysis* 67 (2): 105–111.

Clancy, Sean. 2013. "A Strong Compatibilist Account of Settling." *Inquiry* 56 (6): 653–665.

Clarke, Randolph. 2010. "Intentional Omissions." *Noûs* 44 (1) 158–177.

Davidson, Donald. 1963. "Actions, Reasons and Causes." *Journal of Philosophy* 60 (23): 685–700.

Davidson, Donald. 1973. "Freedom to Act." In *Essays on Freedom of Action*, ed. Ted Honderich, 137–156. London: Routledge and Kegan Paul.

Finch, Alicia. 2013. "On Behalf of the Consequence Argument: Time, Modality, and the Nature of Free Action." *Philosophical Studies* 163 (1): 151–171.

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (3): 829–839.

Frankfurt, Harry. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Franklin, Christopher E. 2011. "Farewell to the Luck (and *Mind*) Argument." *Philosophical Studies* 156 (2): 199–230.

Franklin, Christopher E. 2014. Event-Causal Libertarianism, Functional Reduction, and the Disappearing Agent Argument. *Philosophical Studies* 170 (3): 413–432.

Goetz, Stewart. 2008. *Freedom, Teleology, and Evil*. London: Continuum.

Goldman, Alvin. 1970. *A Theory of Human Action*. Englewood Cliffs: Prentice Hall.

Griffith, Meghan. "Why Agent-Caused Actions Are Not Lucky." *American Philosophical Quarterly* 47 (1): 43–56.

Hornsby, Jennifer. 1980. *Actions*. London: Routledge and Kegan Paul.

Hornsby, Jennifer. 2004. "Agency and Actions." In *Agency and Action*, ed. John Hyman and Helen Steward, 1–23. Cambridge: Cambridge University Press.

Huemer, Michael and Ben Kovitz. 2003. "Causation as Simultaneous and Continuous." *The Philosophical Quarterly* 53 (213): 556–565.

Loss, Roberto. 2009. "Free Will and the Necessity of the Present." *Analysis* 69 (1): 63–69.

Loss, Roberto. 2010. "Fatalism and the Necessity of the Present: a Reply to Campbell." *Analysis* 70 (1): 76–78.

Lowe, E.J. 2006. *Personal Agency*. New York: Oxford University Press.

Lycan, William, G. 1997. *Consciousness*. Cambridge, MA: MIT Press.

Mele, Alfred R. 1992. "Acting for Reasons and Acting Intentionally." *Pacific Philosophical Quarterly* 73 (4): 355–374.

Mele, Alfred R. 1995. *Autonomous Agents*. New York: Oxford University Press.

Mele, Alfred R. 2006. *Free Will and Luck*. New York: Oxford University Press.

Pereboom, Derk. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

Scanlon, Thomas. 2013. "Giving Desert its Due." *Philosophical Explorations* 16 (5): 101–116.

Schlosser, Markus E. 2014. "The Luck Argument against Event-Causal Libertarianism: It Is Here To Stay." *Philosophical Studies* 167 (2): 375–385.

Shabo, Seth. 2013. "Free Will and Mystery: Looking Past the *Mind* Argument." *Philosophical Studies* 162 (2): 291–307.

Shabo, Seth. 2014. "It Wasn't Up To Jones: Unavoidable Actions and Intensional Contexts in Frankfurt Examples." *Philosophical Studies* 169 (3): 379–399.

Steward, Helen. 2009. "Sub-intentional Actions and the Over-mentalization of Agency." In *New Essays on the Explanation of Action*, ed. Constantine Sandis, 295–312. New York: Palgrave Macmillan.

Steward, Helen. 2012. *A Metaphysics for Freedom*. New York: Oxford University Press.

Taylor, Richard. 1966. *Action and Purpose*. New Jersey: Prentice Hall.

van Inwagen, Peter. 1983. *An Essay on Free Will*. New York: Oxford University Press.

Velleman, David J. 2000. *The Possibility of Practical Reason*. New York: Oxford UniversityPress.

Journal of Cognition and Neuroethics

Lessons from Angelology

Edina Eszenyi
University of Kent

Biography

I hold a PhD in Medieval and Early Modern History from the University of Kent in Canterbury. My research field, angelology, and my background in art history have effortlessly led me to Rome, where I am currently based. I work for the Rome Art Program as Lecturer in Art History, and I also steer the art history section of the Program's online blog: www.romeartprogram.org.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Eszenyi, Edina. 2015. "Lessons from Angelology." *Journal of Cognition and Neuroethics* 3 (1): 157–173.

Lessons from Angelology

Edina Eszenyi

Abstract
The article juxtaposes human Free Will with its angelic counterpart through the examination of the c. 1587 *Angelorum et daemonum nomina et attributa*… (Los Angeles, Getty Research Institute MS 86-A866) of the Italian Vincenzo Cicogna (1519? – after 1596). This Catholic reformer author argued for the existence of human Free Will as a negative capacity, demonstrated by the loss of the angelic Free Will following the Fall of the Rebel Angels. In the overall context of the work, his arguments pronounced a wider call for renewal within the Catholic Church, which nevertheless did not resonate with the Inquisition.

Keywords
Angelology, Demonology, Catholic reform, Vincenzo Cicogna

Humans are not generally believed to be the only creatures endowed with Free Will. Angelology, a field where the lack of clear doctrines gives comfortable space for alternative approaches, also recognizes Free Will as one of its central components. Most research into the religious context of Free Will concentrates on the ideas of prominent theologians, however, with the evidence demonstrating how institutionalized ideas failed (and fail) to reach everyday people lurking in the background. Juxtaposing the interpretation of human Free Will with ideas about its angelic counterpart highlights these aspects, while the historical examination of non-mainstream works on angelology brings to surface evidence on how the popular image of angels differed from the image outlined by religious authorities. This alternative reception of theological dogmas in non-mainstream works of literary history is well represented in a manuscript in the collections of the Getty Research Institute in Los Angeles (GRI MS 86-A866).[1]

The elaborate title of the 170 folio Latin manuscript translates as *On the names of angels and demons as found in the Divine Scriptures and explained by the Fathers, dedicated to the illustrious reverend Giulio Antonio Santori, the highest cardinal of Santa Severina, and on the Ecclesiastical Hierarchy* ("Angelorvm et daemonvm nomina et

1. I thank the organisers and participants of the 2014 Free Will conference of the Center for Cognition and Neuroethics for improving, with their questions and comments, the final version of the paper. I also thank the Getty Research Institute for their Library Research Grant and for making the manuscript available online for my PhD research project in the Internet Archive Online Library, accessed 20 December, 2014, http://www.archive.org/details/angelorvmetdaemo00cico.

attribvta passim in divinis scriptvris contenta ad patrvm sententiam explicata ad Illvstriss. et Reverendiss. Ivlivm Antonium Sanctorivm Cardinalem Sanctae Severinae amplissimvm et de Ecclesiastica Hierarchia"). The manuscript divides into two main sections. The first, Lexicon section is a collection of particular and metaphorical references to angels and demons, listing altogether 100 angel and 123 demon keywords in alphabetical order. The keyword selection is based mostly, but not exclusively, on the Bible. The second main section is a treatise which draws a parallel between the angelic and the ecclesiastical hierarchies, and argues that the Church fails to follow the heavenly example of angels. The work as a whole was dedicated to the powerful Cardinal Giulio Antonio Santori (1532–1602), surprisingly enough as the concluding treatise compares cardinal bishops, among them the dedicatee, to Cherubs, identified in the Lexicon as the original order of fallen angels.

The author signed his work as Vincentius Ciconia, and is identifiable with an ecclesiastical author known as Vincenzo Cicogna who lived and worked in Verona, Italy. Cicogna was born in 1519,[2] when Verona, within the jurisdiction of Venice, was going through a defining ecclesiastical reform process under the local bishop Gian Matteo Giberti (1495–1543).[3] Bishop Giberti had been a talented diplomat of the Holy See, who moved to his diocese after the Sack of Rome in 1527. The conditions upon his arrival inspired him to eliminate the obstacles hindering his clergy at becoming proper guides of the population. This called for an overarching regulation from everyday routine to professional aspects. The bishop approached the implementation of the necessary moral and disciplinary changes in a systematic way: first he secured the authority for interventions, then he gained firsthand experience of the local situations by pastoral visits, and finally secured the changes by written regulations. Throughout the construction and implementation of the reforms, Giberti cooperated with a group of learned ecclesiasts, who recognized a need for higher-level reforms and also voiced concerns over the Church's own ability of renewal. Vincenzo Cicogna was a member of the bishop's specially trained clergy.[4]

2. Archivio di Stato di Verona, *Anagrafi Comune* 1210; Archivio di Stato di Verona, *Antico Archivio del Comune*, Anagrafe 1215.

3. For a general bibliography see Turchini 2000.

4. Previous research has not yet clarified the exact status of this specially trained clergy around the bishop. Scholarship simultaneously reports a special school of the bishop (*schola Accolythorum*) and a group of learned intellectuals living in Giberti's household (*familiares*) without clearly defining the differences and similarities between the two (Prosperi 1969, 226-234; Cervato 1989, 39-55; Eszenyi 2014, 58-60).

Vincenzo became the first ecclesiastical member of a popular local painter dynasty of Greek immigrants, probably due to the reformer bishop's friendship with his father.[5] The adventurous priest was a member of Giberti's intellectual circle with all the benefits and risks the membership attracted. The bishop's authority granted him professional networks reaching as high as Charles Borromeo, whom Cicogna assisted in pastoral visits in 1564, but also secured his local position as rector of the San Zeno in Oratory monastery between 1544 and 1566 (Rognini 2004, 10; Tacchella 1979, 128–129, 132). The influence of the Giberti circle nevertheless also resulted in Cicogna's first direct encounters with the Inquisition in an 1550 series of Verona trials, when his preaching was found to be 'a fountain of heresy' (Conforti 2004, 104; Tacchella 1979, 128–129).

The mid-century was also the period when Cicogna started to publish, commonly dedicating his works to high level ecclesiasts. His first two works were sermon collections with unquestioned orthodoxy, perhaps one of the reasons for their publications being exactly Cicogna's desire to clarify himself from the early accusations of heresy. *Sermones 7* (Venice, 1556) was a collection of seven sermons on the Eucharist, dedicated to Aloysio Lipomano, bishop of Verona. These early sermons were republished and accompanied by six new Passion sermons in *Sermones* (Venice: Andrea Arrivabene, 1562), dedicated this time to Cardinal Marcantonio Da Mula (Amulio). The *Oratio in Bernardi Naugerii cardin*[alis] *amplissimi et episcopi veronen*[sis] *aduentu* (Venice: Iordani Zileti, 1564) was an oratory speech given by Cicogna when Cardinal Bernardo Navagero paid a visit to Verona.[6]

Cicogna's problems with the Inquisition nevertheless persisted. His best-known theological work, the 1567 *Enarrationes in psalmos*, was '*nisi corrigantur*' prohibited and included in the 1580, 1583, and 1596 Indexes of Prohibited Books. It was a commentary of Psalms 118-133 (119–134 today), accompanied by meditations on letters of the Hebrew alphabet, and dedicated to Pius V. The next in line of his censored works is lost, but the Archives of the Congregation of the Doctrine of the Faith, the congregation formerly administering the Inquisition, preserved its dedication. The work was entitled *Thesaurus d‹ivina› oracula et attributa continens* (*Collection of divine prophecies and*

5. On the family see Simeoni 1907; Da Re 1913; Brenzoni 1958; Brenzoni 1972; Guzzo 1996; Varanini 1996; Eszenyi 2014, 22-29.

6. Another published speech, composed at the 1565 death of the same cardinal, is mentioned in Jacopo Vallarsi and Pisrantonio Berno, *Verona Illustrata parte seconda* (1731, 422). The same source also claims that the *Enarrationes in psalmos* appeared in print already in 1556, but I found no confirmation of this early work's claims in other sources.

attributes).[7] The dedication addressed pope Gregory XIII, consequently its composition is datable to 1572-1585. An undated letter attached to the *Thesaurus*' dedication in the Archives prohibited Cicogna from publishing or even composing anything related theology in the future,[8] and its author was imprisoned in Rome for six months in 1573 for reasons currently unknown to research (Da Re 1913, 119; Guzzo 1996, n. 40).

Nevertheless, the *Thesaurus* is probably identical with two volumes on divine names and prophecies Cicogna had sent to Cardinal Santori, the *Angelorum*'s dedicatee, for publication in the year prior to composing the *Angelorum*'s dedication, as the latter text reveals.[9] Cardinal Giulio Antonio Santori (1532–1602), Cardinal of Santa Severina from 1579 until his death, was an outstanding personality of his times: "he influenced all the affairs of the Church in the last third of the 16th century as few other members of the Roman Curia" (Santori et alia 1966, 5). He acted as personal consultant of several popes and was, in 1592, himself a candidate for papacy. Besides being an advocate for the union of the Eastern and Western Churches he was also a productive man of letters, who composed numerous liturgical, historical and canon law works, as well as personal writings. On the request of Pope Paul V, the cardinal also composed a sacerdotale in 1586, a work which later provided the foundations of the current Roman Ritual. Cardinal Santori is yet best known as Prefect of the Sacred Congregation of the Roman and Universal Inquisition, practically Italy's most powerful Grand Inquisitor in this position. He participated in the heresy processes of historical characters such as Giordano Bruno, Tommaso Campanella, Cardinal Giovanni Morone, or Henry of Navarre – and ordinary people such as Carlo Ginzburg's Menocchio. This position within the Inquisition naturally granted Cardinal Santori the overview of the Index of Prohibited Books, which could well have raised the interest of a persecuted author such as Vincenzo Cicogna (Ginzburg 1982, 127–128; see Ricci 2002 for further bibliography on Cardinal Santori).

The *Angelorum* managed to get attention from its dedicatee but apparently was not fully, or perhaps at all, welcomed. The Archives of the Congregation of the Faith register

7. Archivio della Congregazione per la Dottrina delle Fede, Index *Protocolli* G, fols 306^{r}–317^{r}.

8. I thank Dr. Barbara Bombi at the University of Kent Canterbury for the transcription and translation of the text.

9. "Dei Opt. Max nomina, Ill‹ustrissim›o Paesul, et attributa, passim in sacris literis contenta in unum redegi volumen, per tres divinas personas atributa, ad sanctissimorum Patrum explicate sententiam. Quod volume anno superiori tuo iussu Romam, cum altero volumine oraculam ad Christi fidem spectantia complectente, transmisi: ut censua et iudicio Sedis Apostolicae…" (GRI MS 86-A866, Fol. 1^{r}).

a document containing Inquisitorial notes about it, by all probability corrections.[10] They are attributed to bishop Federicus Metius, a censor referred to as a *familiaris* of Cardinal Santori in numerous lists of censors (*consultores*) of the Index, normally among primary censors.[11] The exact nature of the inquisitorial corrections remains a challenge for further research until the now lost censorship document hopefully resurfaces one day. Til then, the *Angelorum*'s dedication, Cicogna's literary oeuvre and biographical data point towards the year 1587 as a likely date for the *Angelorum*'s completion, and make it likely to be Cicogna's last work.

Vincenzo Cicogna, Bishop Giberti, and Cardinal Santori were the most important characters in the events leading to the creation of an unusual angel lexicon in a world shaped by the vast turmoil of major religious upheavals. The *Angelorum* is more than theological in nature. More than simply pointing out problems with the Church in the language of angelology, its author took a constructive approach by arguing for the universal nature of Christianity, which he tried to demonstrate by highlighting its understated harmony with pre-Christian philosophical systems. How is Cicogna's angelology relevant for the study of human Free Will? He argued for the existence of human Free Will as a negative capacity, demonstrated by the loss of the angelic Free Will following the Fall of the Rebel Angels.

Cicogna's *Angelorum* recounts that all angels were created good and endowed with Free Will. Lucifer among them, an outstandingly beautiful Cherub on top of the angelic hierarchy, had a unique relationship with God, but fell from this status when he committed a sin, specified by Cicogna varyingly as dissatisfaction, pride, and the misuse of power. The largest part of angels decided to join Lucifer, as a result of which they had to be cast out of Heaven and be separated from the good angels. Besides the physical separation, they are now also separated by a different name: demons. Following the

10. "*Vincentii ciconii de nominibus Angelorum et demonum p‹er› federicu‹m› Metiu‹m› f‹ol›. 567*" Archivio della Congregatione per la Dottrina della Fede, Index *Protocolli* D, Fol. 3^{v}. The volume contains documents dated after 1575.

11. Archivio della Congregazione per la Dottrina della Fede, Index *Protocolli* I, fols 359^{v} and 361^{v} use the term "*familiaris*", fols 362^{r} and 360^{r} list him in the first class of inquisitors, Fol. 366^{r} specifies that Metui was charged with censorship of books (*quibus assignati sunt Libri ad Censurandum*), Fol. 373^{r} lists him in a list of *consultores* without further specification. The volume contains documents dated to 27 April 1573 – 28 June 1593.

events, both angels and demons were confirmed in their chosen good or bad natures and from that time on, they are not able to and neither do they will to act otherwise.[12]

Simple as the story might seem to be, there is a lot to explain about the Fall of the Angels in a work on Biblical interpretation as this widespread tradition has no undisputable Scriptural base. Despite a number of passages understandable as possible references to the story, the Fall of the Angels has no clear phrasing in the Scriptures and the Free Will of angels is not associated with any possible coverage of the story. Tradition was clearly the heavier component in its increasing popularity, probably influenced by apocryphal writings as much as its inclusion in the Legenda Aurea, its eyecatching compositions in art, and less but not least, by the high number of theologians who could not resist the allure of interpreting these mysterious events in their writings (Eszenyi 2015).

While Cicogna's approach owes much to prominent medieval theologians, his interpretation is anything but common. Thomas Aquinas' *Summa Theologiae*, a work cited in Cicogna's *Angelorum* on several occasions, is arguably the most influential work discussing the question prior to Cicogna. Aquinas argued that angels have will with a stronger natural tendency towards good than man but he described no connection between the angelic Will and the Fall of the Angels (Aquinas 1920). Divine confirmation nevertheless plays an essential part in Cicogna's system, as it is explained under the keyword *Caeli* in the Lexicon section of the *Angelorum* (Fol. 23$^{r\text{-}v}$). At the moment of creation, the sky is called *caelum* in Genesis 1:1, but it is referred to as 'firmament', *firmamentum* in Latin, with a sudden change from Genesis 1:6. Cicogna suggests that the switch expresses the 'firmness' good angels gained from the divine confirmation following the Fall of the Rebel Angels, because the sky symbolically refers to Heaven, their dwelling place.[13]

12. Characteristically of the *Angelorum*'s argumentation technique, the concept of the Fall of the Angels does not stand as an isolated argument but recurs in numerous Lexicon entries, some among which shed more light on the author's views by turning the theme into the entry's primary argument. The most informative Lexicon entries are *Astra Matutina* (Fol. 18$^{r\text{-}v}$), the double entry on Cherubs (*Cherubim* on fols 25^{r}-26^{r} and *Cherub* on Fol. 110$^{r\text{-}v}$), *Drachmae* (fols 36^{v}-37^{r}), *Lucifer* (double entry on fols 55^{r} and 134$^{r\text{-}v}$), *Lapides* (fols 51^{v}-52^{r}) *Michael* (fols 56^{v}-57^{r}) *Signaculum similitudinis* (double entry on fols 77^{v}-78^{r} and 198^{v}-199^{r}), *Stellae* (fols 79^{v}-80^{r}), *Draco* (Fol. 116^{v}), *Fulgur* (Fol. 123$^{r\text{-}v}$), and *Principium* (fols 151^{v}-152^{r}).

13. "Quod bene apud Moisem ipsa de mundi historia verba testantur, cum et prius celum factum dicitur: et hoc idem postmodum firmamentum vocatur: Quia videlicet natura Angelica et prius subtilis est in superioribus condita: et post, ne potuisset unqua‹m› cadere, mirabilius confirmata" (GRI MS 86-A866 Fol. 23^{r}).

Let's now compare the situation of the two groups and see why Cicogna's good angels largely profit from the divine confirmation. Among a range of fascinating metaphors, he also compared angels to electrum, the natural alloy of gold and silver in the entry with the keyword *Electrum* (fols 39^{v}-40^{r}). As the electrum is not fully gold but mixed with a material second only to gold, angels are not divine, but they are very close and very similar (*quamsimillimi*) to the divine,[14] and one manifestation of their similarity is the inability to oppose the divine will. Divine confirmation makes angels the only creatures who never oppose the divine will: demons are never obedient, humans are hindered by their Free Will, but the will of angels equals the will of God. The Word of God is incomprehensible for angels as much as for humans, yet angels gain all knowledge and diligence through divine revelation originating from the contemplation of the face of God. This direct observation of the divine teaches angels everything that they later pass on to other creatures, with the intention of making them obedient to God by perfectly fulfilling the divine will, which equals their own. In short, angels lost the ability to choose between good and bad as a result of the divine confirmation, which is in a sharp contrast with their pre-fall state. They are content, however, as neither do they desire the freedom of choice anymore.[15]

Let us now have a look at fallen angels or demons, whom Cicogna also described with abundant metaphors and allegories. How does their situation compare with that of the good angels? Demons remained highly intelligent after the fall but use this intelligence to separate people from God, says Cicogna. They are irreconcilable enemies, showing benevolence but only disseminating heresy among people like weeds among the wheat in the Parable of the Weeds in Matthew 13:24–30.[16] Furthermore, demons

14. "Nam cum auro et argento nihil inter metala sit praeciosius: ita Angeli nobilitate caeteras superant creaturas. Non sunt purum aurum: quia neq‹ue› Deus neq‹ue› ex Dei substantia su‹n›t: sed Deo proximiores, et quamsimillimi sunt… Cum itaq‹ue› Angeli Electro comparantur, declaratur illos splendor praefulgare, et spectantibus non mediocrem praebere consolationem" (GRI MS 86-A866 fols 39^{v}-40^{r}).

15. "Hoc unum est omnium Angelorum opus et exercitiu‹m›, ut benedicant Domino omni tempore, et voluntatem eius faciant. … Dei enim voluntas aeterna est, et incomprehensibilis non solum hominibus sed etiam Angelis ipsis: de qua quicquid eis a Deo est revelatum, explere quidem possunt, sed eam, prout est, totam capere nequeunt. … Angeli itaq‹ue› semper vide‹n›t faciem patris, qui nihil facit, quod sanctis suis non revelet. Intelligentie itaq‹ue› oculis, Dei voluntatem illo revelante… Soli Daemones imperfracte voluntati illius revelatae resistant… Homines etiam cum libero arbitro agant, quandoq‹ue› Dei voluntati apertae repugnant, quod ille prohibeat volentes…" (GRI MS 86-A866 Fol. 44^{r}).

16. "Demonis autem nomen etsi mentem sapientem significet, pro malo tamen spiritu usurpetur, qui sapiens est, ut faciat malum: quod potius est desipere, quam sapere" (GRI MS 86-A866 Fol. 3^{v}). "Inimicus homo,

also have a quite particular task: they are occasionally sent by God to announce bad news and to complicate the lives of sinful people. Cicogna presents demons as servants of the divine justice, who execute divine justice on disobedient people with corrupt souls. Scriptural examples include the angel who released the plague upon Israel in 2 Samuel 24:15-25, or the angel who caused the Ten Plagues of Egypt in Exodus 5–12.[17]

The existence of one particular angel whose duty is the execution of divine punishments was not unknown to medieval thinking. This angel was often referred to as *the* Angel of the Lord as opposed to *an* angel of the Lord, called *Angelus Domini* by his Latin name. Cicogna says the name *Angelus* expresses that this angel is on a mission, however dark it might be, as angels are primarily divine messengers. The name the Angel of the Lord is a reminder that even though demons sinned on their own will, they were still created by God, and their dark powers are strictly limited to what divine providence allows.[18]

Cicogna nevertheless seems to hesitate: at one point he suggests that not only one angel is charged with the task but it is a collective responsibility of demons, whereas in other parts of the text he refers to this angel in the singular. If Cicogna had one particular angel in mind it must have been Lucifer as in one point he identifies this mysterious punishing angel with the fallen dragon serpent of Revelations 12:7–9, cast out of Heaven after a battle with good angels under the leadership of Archangel Michael. Identifying the Angelus Domini with Lucifer could possibly stand as an individual idea by Cicogna, inasmuch as I have not found examples of the same idea in medieval angelology yet. The text itself does not clearly reveal if the singular-plural inconsistency was intentional

---

qui scilicet per hominem malum Zizania idest, haereses disseminat in agro Domini appellatus a Christo Diabolus: […] publicus et privates sit inimicus irreconciliabilis: qui tunc etiam inimicitias great, cum amicitiam et benevolentiam praeseferat…" (GRI MS 86-A866 Fol. 129^{r}). Cicogna mistakenly refers to Matthew 14 instead of Matthew 13 on the margin.

17. "Nam ut bona per bonos Angelos: ita mala per malos confert Deus: eaq‹am› secundum sibi constitutum modum et mensuram: Etsi enim inuiti et malo animo Dei iustitiam administrant, non tamen sibi praescriptum mensura‹m› excedere queunt..." (GRI MS 86-A866 Fol. 141^{r}).

18. "Dicitur ergo Angelus Dei et a Deo: suo proprio vitio statim factus est malus. Cum itaque creatura Dei sit, Angelus Dei est et dicitur. Et cum idem etiam invitus Dei subiectus sit, neque quicquam possit, nisi id sibi per Deum liceat, a DEO esse, et mitti dicitur. Unde etiam Angeli nomen retinuit, quod Missus significat, cum nomine etiam quaedam alia sibi cum bonis Angelis communia retinens: Est enim spiritus sicut et illi, et Dei administer licet in malis, sicut et illi in bonis. Habet vires sicut et illi, sed illis ipse uti non potest, nisi id sibi a Deo per Angelos bonos sit permissum" (GRI MS 86-A866 fols 97^{r-v}). For the Angelus Domini problem see for example Fossum 1985, Deutsch 1999; White 1999; with further bibliography.

on part of the author. Cicogna himself could also have been simply undecided about the question, which had never gone undisputed in medieval angelology. By no way does it lessen the value of the peculiarity of his arguments though for punishing angels with limited divine powers, especially when it comes to the question of Free Will. This makes either Lucifer or all demons unwilling servants of the Lord.[19]

Yet Cicogna attributes different powers to demons when they are and when they are not in divine service. With a deceitful nature, they have a tendency to approach people even if they are not sent by anybody, only to mislead, by lies, those who don't exercise proper care at the discernment of spirits. Cicogna says the power of both angels and demons is insuperable for humans when the spiritual messengers are executing the divine will. Nevertheless demons approach us with temptations only in the majority of cases, and in these cases they can and should be overcome.[20]

Cicogna does not fail to notice that the angel refused worship and directed it to God instead in a conversation with the prophet John in Revelations 19:10 and 22:8-9, with the explanation that he is but a fellow servant of the prophet.[21] Angels, demons, and humans are all divine servants in Cicogna's opinion, but where is the difference between these three types of servants? Cicogna answered this question by embracing the theory of a cosmological tableau of cosmic order, the ancient Greek idea of the Great Chain of

19. "Angeli itaque appellantur Diaboli: quod ipsi quoque a Deo mittantur, tanquam furoris indignationis tribulationis et gladii sui administri contra rebelles et peccatores: Ut enim bona omnia per bonos, ita mala per malos Angelos confert Deus: Et ut boni in bonorum administratione bene operari, ita isti in malorum immissione peccare dicuntur" (GRI MS 86-A866 Fol. 97^{r}). "Dicitur ergo Angelus Dei et a Deo: suo proprio vitio statim factus est malus. Cum itaque creatura Dei sit, Angelus Dei est et dicitur. Et cum idem etiam invitus Dei subiectus sit, neque quicquam possit, nisi id sibi per Deum liceat, a DEO esse, et mitti dicitur" (Fol. 97^{r-v}). These latter remarks are added to the end of the entry by a Second Hand, who often made additions and corrections to the main text – supposedly the author himself making additions to the secretarial handwriting of the First Hand.

20. "In hoc autem differunt spiritus mali a bonis, quod illi et iussi et non iussi temere nunciant et agunt. Isti vero non nisi iusti agunt et nunciant: Illi spiritus sunt mendaces, et mendacia loquuntur" (GRI MS 86-A866 Fol. 12^{v}). "...neque sit potestas super terram, quæ ei [demons] resist[ere] queat: ... sed quamvis spiritus iste pervalidus sit, dicitur tamen et ipse spiritus Domini *vel* a Domnio egressus: quod nihil omnino possit, invito Deo; cuius ministerio utitur, cum iustitia illius supplicia de peccatoribus exigit: propterea Spiritus tempestatis et furoris Domini appellatur: ut enim bona per bonos ita mala per malos Angelos et spiritus immittit Deus" (GRI MS 86-A866, Fol. 155^{v}).

21. "Ea erat Ioannis humilitas, ut Angelum honore praevenire, et venerari voluisset: Sed tanta etiam est Angloru‹m› charitas, ut quos sibi a Deo coaequatos videant, inferiores sibi esse non permittant, et illos suos conserves appellent quod eundem Dominum habeant…" (GRI MS 86-A866 Fol. 30^{r}).

Being in full revival among Early Modern magicians of his time.[22] The opening lines of his *Angelorum* take as a starting point that divine wisdom and providence gave a hierarchy to creatures, where angels occupy the highest position above man, animals, and of course demons.[23] The idea of predestination also surfaces in Cicogna's angelology when he adds that angels will spread divine love and mercy to selected people, and the selected will be confirmed just like the obedient angels were confirmed.[24]

Why this emphasis on hierarchy, in Heaven and Earth, before and after death? Examining the contents of the Lexicon section in the wider context of the *Angelorum*'s closing treatise offers an explanation. Cicogna argues for the leading and exemplary role of angels throughout the work, supporting his arguments with Scriptural passages such as the angel leading the Israelites from Egypt to the Promised Land in Exodus 23. He argues that their close proximity to God grants angels excellence over other creatures: second to God only they stand incorrupt, humble, and always in agreement with one another.[25] They are almost omnipotent, with powers dependent on God only. Their greatest virtue is the ability to pass on this power to people, and they are good spiritual leaders because they provide examples.[26]

In context of the manuscript's closing treatise drawing a parallel between angels and the clergy, this point echoes bishop Giberti's reform ideas about the clergy standing as an example for the people. Giberti re-organized religious life with the aim of creating a clergy that functions as an example for the people – similarly to the way the clergy was supposed to mirror angels in Cicogna's *Angelorum*. The idea of the preacher whose main duty is to teach and inspire was also outlined in a manual bishop Giberti's press

22. On Early Modern angelology see Marshall and Walsham 2006; Bailey 2007; Fanger 1998; Keith, 1997.

23. "Divinæ sapientiæ et providentiæ congruum esse videbatur, ut cum creaturas condere statuisset, eas /ut scribet Sapiens/ in pondere numero et mensura crearet. Propterea quamvis */ut idem docet/* omnia simul creaverit, eas tamen ordinati condidit: et ex his alias præstantiss‹im›as, alias mediocres, alias vero infimas esse voluit. Primum et præstantiss‹im›um locum apud se Angelos, medium vero hominem, infimum habere voluit belluas" (GRI MS 86-A866 Fol. 3r).

24. "Quod Angeli Dei dilectionem et misericordiam erga se et caeteras creaturas perpetuis laudibus efferant. Ab aeterno siquidem et ante tempora secularia electos suos tum ex Angelis tum ex hominibus dilexit… Sed ut Dei iustitia poscebat, ut Angeli rebelles ab accepta gratia deiicere‹n›tur: ita Dei dilectio et misericordia voluit, ut qui in veritate stetissent, in ea ita confirmati et stabilitia essent…" (GRI MS 86-A866 Fol. 49^{v}).

25. "Excelsi autem appellati sunt, quod dignitate et excellentia virtutu‹m›, omnes creaturas celestes, terrestres, et infernales superent, et Deo proximiores sint… nulla tamen est inter illos discordia…" (Fol. 42^{r}).

26. "Est enim Angelus Dei dilectionis et misericordiae utuum exemplar: quandoquide‹m› in illis, nobis ardere et contemplari licet, quae et quanta sit Dei charitas et dilectio…" (GRI MS 86-A866 Fol. 49^{v}).

printed in 1544, entitled *For the Preaching Fathers* (*Per li padri predicatori*). This detailed practical guide for the evangelization instructed priests to teach the population not only with preaching but also once they leave the pulpit, persuading people with the example of their very own lives. Giberti saw the ideal clergy as living an extremely severe and elevated life, which the bishop himself also practiced in his own household (Segala 1989; Prosperi 1969, 201, 215, 231, 251–252, 261–262, 180–182).

Perhaps inspired by the impressive diplomatic careers of Bishop Giberti and Cardinal Santori, Cicogna raised this idea to a political level and extended spiritual guidance to secular governance. Angels command people as people command animals, and God commands angels as angels command people, he said, with the divine origins of angelic power granting its legitimacy. Secular forms of leadership are rather problematic, their only correct form is the one guided by angels. With reference to Gregory the Great, Cicogna interpreted the expression 'kings and counsellors' in Job 3:13–15 as a metaphor expressing that God rules this world with the help of kings acting upon the spiritual counsel of angels. In light of the closing treatise comparing angels to members of the clergy, angels advising kings are not difficult to understand as a call for clerical advisors by the sides of kings.[27]

Most of the Lexicon entries commenting upon angelic leadership are doubled in Cicogna's *Angelorum*, the *De Demoniis* section providing counterexamples of the ideals outlined in the *De Angelis* section of the work. Demonic activities explain the negative potentials of government. Tyrants are the opponents of kings, representatives of bad government influenced by demons, which explains the title 'king' being attributed to the Devil in Job 41:34. Cicogna states that non-believers, just like devils, have a strong desire to exercise a restrictive authority over others, therefore it is proper to call their leader a king. Yet this is a king who reigns insufficiently, exercising tyranny. Cicogna stresses that the Job passage associates tyrants and non-believers with pride and reminds his reader

27. "Quia /ait [Gregory the Great]/ cunctorum conditor omnia per semet ipsum tenet: et tamen ad distinguendum pulchræ universitatis ordinem alia aliis dispensantibus regit, non immerito Reges Angelorum spiritus accipimus: qui quo omnium authori familiarius serviunt, eo subiecta potius regunt... Qui bene etiam consules vocantur, quia spiritali Rei p‹er› bene consulunt, dum nos sibi socios adiungunt... Bene Consules vocantur, quia dum ipsis nunciantibus voluntatem conditoris agnoscimus... quorum omnium saluti cum consulant, Deum consulentes, merito Reges et Consules, seu Consiliarii /ut vox chaldaica sonat/ sunt appellati: idque præsertim, quod non solum Deus sua consilia illis committit, sed quod nos saluberrimis consiliis monent, quibus nostram ipsorum salutem consequi, et exitium cavere possimus: si ergo recte rebus nostris est consultum, id Angelis acceptum ferre debemus, a quibus est omne bonum consilium" (GRI MS 86-A866 Fol. 73^{r}). On the relationship between clergy and monarchs see for example Hay 1977.

that the king of non-believers, the devil, fell of the same sin at the Fall of the Angels. By associating Satan with the sin of pride that caused his fall, Cicogna closes the circle of his argumentation and returns to the Fall of the Angels, the popular tradition building a bridge between his angels and demons.[28]

Cicogna, in short, argued for the existence of human free will as a negative capacity, demonstrated by the loss of the angelic free will. Angels in his understanding were all created good and endowed with free will, but the largest part of them intentionally chose to turn bad. As good angels were confirmed in their goodness by divine grace after the Fall of the Rebel Angels, so were rebel angels, now demons, confirmed in their malicious nature. Contrary to their pre-fall state, their will is now limited to what is in accordance with their predefined good or bad nature. Both groups lost their ability to choose between good and bad, but good angels largely profit from this loss by earning a place second only to God in the existential Hierarchy of Beings, due to their inability to sin. Demons can never agree to the divine will, yet they are paradoxically forced to serve it as executors of divine punishments. Humans occupy an in-between position between demons and angels as Free Will hinders their ability to act upon the divine will. Consequently humans should be obedient to angels no less than to God, which is understandable as a call for obedience to the clergy in light of the closing treatise of Cicogna's *Angelorum*.

Cicogna's angelology challenged the traditional positive evaluation of human Free Will in lines that rhymed with the ideology of his own conservative Catholic reformer background. He did not necessarily correspond to the mainstream approaches and fashionable intellectual trends of his age or ours, but demonstrated the potentials within a yet unclarified and somewhat obscure tradition. His theory about Lucifer's possible identification with the Angelus Domini as a punishing angel could be a novelty within the field of angelology itself. One can only wonder what impact such a highly educated,

28. "Diabolus etiam potens est… in dolo in peccato: quae quidem potestas Tyrannidis potius nomen habet… si enim illi nos assenserimus, potens efficitur, si resistim‹us› fugit, et imbecillis efficitur… At Angeli potentes sunt virtute: quia illoru‹m› virtues et fortitudo est Deus… Verbum Dei est voluntas illius" (GRI MS 86-A866 Fol. 68^{v}). "Rex super universos filios superbiæ a Iobo: a Salomone vero Rex magnus obsidens civitatem parvam, in qua sit pauper eam ad obsidione liberans, appellatur Satan… Hic obsedit civitatem parvam, in qua est pauper: Ipsa est Ecclesia, in qua est Christus, quam divexare quidem, sed capere aut destruere non potest… Dicitur etiam Diabolus Rex super omnes filios superbiæ; quod ille unus superbia sua omnes superbissimos ea celat et superet: cum non solum omnibus creaturis, sed ipsi etiam creatori se prætulerit: et perpetuo studeat suæ superbia‹e› habere imitatores: Superbiam, ut ipse, ita et homines matrem et altricem habere vult" (GRI MS 86-A866 fols 153^{v}-154^{r}).

well-networked, successful but controversial author could have made, had it not been for the Inquisition's vigilance, which still preserved Cicogna's angelology as a captive and captivating representative of the intellectual diversity of the 1500s.

References

Primary sources

Aquinas, St. Thomas. 1920. *The Summa Theologica of St. Thomas Aquinas*. Second and Revised Edition. http://www.newadvent.org/summa/1059.htm.

Archivio della Congregatione per la Dottrina della Fede, Index *Protocolli* D.

Archivio della Congregatione per la Dottrina della Fede, Index *Protocolli* I.

Archivio di Stato di Verona, *Anagrafi Comune*. 1210.

Archivio di Stato di Verona, *Antico Archivio del Comune*, Anagrafe. 1215.

Cicogna, Vincenzo. 1585. *Angelorum et daemonum nomina et attributa*. Los Angeles: Getty Research Institute MS 86-A866. In *Internet Archive Online Library*, http://www.archive.org/details/angelorvmetdaemo00cico.

Krajcar, John. 1966. *Cardinal Giulio Antonio Santoro and the Christian East: Santoro's Audiences and Consistorial Acts*. Orientalia Christiana Analecta, 177. Roma: Pont. Institutum Orientalium Studiorum.

Secondary literature

Bailey, Michael D. 2007. *Magic And Superstition in Europe: A Concise History from Antiquity to the Present*. New York: Rowman and Littlefield.

Brenzoni, Raffaello. 1972. *Dizionario di artisti veneti*. Firenze: L.S. Olschki.

Brenzoni, Raffaello. 1958. "Un fresco del '500 e una tempera della fine del '400.'" In *Miscellanea in honore di Roberto Cessi*, 55–67. Roma: Edizioni di storia e letteratura II.

Cervato, Dario. 1989. "I collaboratori." In *Gian Matteo Giberti Vescovo di Verona 1524-1543,* 39–55. Verona*: Biblioteca Capitolare di Verona*.

Conforti, Giuseppe. 2004. "Villa Del Bene: iconografia e inquietudini religiose nel Cinquecento. Gli affreschi della loggia e dell'Apocalisse." In *Annuario Storico della Valpolicella 2003-2004*, edited by Andrea Brugnoli, 99–119. Verona: Centro di Documentazione per la Storia della Valpolicella.

Da Re, Gaetano. 1913. "I Cicogna dal secolo XVI." *Madonna Verona. Bollettino del Museo Civico di Verona* 7: 109–123.

Deutsch, Nathaniel. 1999. *Guardians of the Gate: Angelic Vice Regency in Late Antiquity*. Leiden: Brill.

Eszenyi, Edina. 2014. "On Perfect and Imperfect Angels: A Catholic Reformer's Angelology from the Late-Sixteenth Century Veneto." PhD dissertation. Kent: University of Kent.

Eszenyi, Edina. 2015. "Thunderbolt. Shaping the image of Lucifer in the Cinquecento Veneto." In *Proceedings of The Marriage of Heaven and Earth: Images and Representations of the Sky in Sacred Space: Sophia Centre for the Study of Cosmology in Culture 12th Annual Conference*.

Fanger, Claire. 1998. *Conjuring Spirits: Texts and Traditions of Medieval Ritual Magic*. University Park: Pennsylvania State University Press.

Fossum, Jarl E. 1985. *The Name of God and the Angel of the Lord: Samaritan and Jewish Concepts of Intermediation and the Origin of Gnosticism*. Tübingen: Mohr Sieback.

Ginzburg, Carlo. 1980. *The Cheese and the Worms: The Cosmos of a Sixteenth-Century Miller*. Translated by John and Anne Tedeschi. Baltimore: Johns hopkins University Press.

Guzzo, Enrico Maria. 1996. "Il palazzo Del Bene di San Zeno in Oratorio in Verona (e le relazioni di Giovanni Battista Del Bene con alcuni artisti veronesi)." In *La famiglia Del Bene di Verona e Rovereto e la villa Del Bene di Volargne, atti della giornata di studio, Rovereto e Volargne 30 settembre 1995,* edited by Gian Maria Varanini, 81–113. Rovereto: Accademia Roveretana degli Agiati.

Hay, Denys. 1977. "The Italian Renaissance and the Clergy of Italy in the Fifteenth Century." In *The Church in Italy in the Fifteenth Century***.** Cambridge: Cambridge University Press.

Keith, Thomas. 1977. *Religion and the Decline of Magic: Studies in Popular Beliefs in Sixteenth and Seventeenth Century.* New York: Oxford University Press.

Marshall, Peter, and Alexandra Walsham. 2006. *Angels in the Early Modern World*. Cambridge: Cambridge University Press.

Prosperi, Adriano. 1969. *Tra evangelismo e controriforma. G. M. Giberti (1495–1543)*. Rome: Edizioni di storia letteratura.

Rognini, Luciano, 2004. *La chiesa di San Zeno in Oratorio. Guida storico-artistica*. Verona.

Segala, Franco. 1989. "Pastore e riformatore." In *Gian Matteo Giberti Vescovo di Verona 1524–1543,* 27–30. Verona*: Biblioteca Capitolare di Verona.*

Simeoni, Luigi. 1907. *Maestro Cigogna (1300–1326)*. Verona: Antonio Gurisatti.

Ricci, Saverio. 2002. *Il Sommo Inquisitore: Giulio Antonio Santori Tra Autobiografia e Storia (1532-1602)*. Roma: Salerno.

Tacchella, Lorenzo. 1979. *Il Processo agli eretici veronesi nel 1550. S. Ignazio di Loyola e Luigi Lippomano.* Brescia: Morcelliana.

Turchini, Angelo. 2000. "Giberti, Gian Matteo." In *Dizionario biografico degli Italiani*, Vol. 54., 623-629. Rome: Istituto della Enciclopedia Italiana Fondata da Giovanni Treccane S. p. A., online in *Dizionario biografico degli Italiani*. http://www.treccani.it/enciclopedia/gian-matteo-giberti_%28Dizionario-Biografico%29/.

Varanini, Gian Maria. 1996. "Il pittore Nicola Crollalanza e gli affreschi di Villa Del Bene (1549)." In *La famiglia Del Bene di Verona e Rovereto e la villa Del Bene di Volargne, atti della giornata di studio, Rovereto e Volargne 30 settembre 1995*, 149–165. Rovereto: Accademia Roveretana degli Agiati.

White, Stephen L. 1999. "Angel of the Lord: Messenger Or Euphemism?" *Tyndale Bulletin* 50 (2): 299-305.

Journal of Cognition and Neuroethics

Exploring the Status of Free Will in a Deterministic World: A Case Study

Catherine Gee
University of Waterloo

Biography

Catherine Gee is a PhD student in philosophy at the University of Waterloo in Waterloo, Ontario. Her primary research interests lie at the intersection of philosophy and psychology, and in philosophy of psychiatry in particular. Issues concerning the proper classification of mental disorders and their implications for treatment are one of the current topics she is working on, in addition to projects in philosophy of mind and philosophy of science.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Gee, Catherine. 2015. "Exploring the Status of Free Will in a Deterministic World: A Case Study." *Journal of Cognition and Neuroethics* 3 (1): 175–194.

Exploring the Status of Free Will in a Deterministic World: A Case Study

Catherine Gee

Abstract
Developments in sciences have been uncovering causal mechanisms which have improved our understanding of human character and behaviour, which seem to be the result of deterministic forces. This leads to questions regarding the status of free will as it is often argued to be incompatible with determinism. Individuals with anorexia nervosa present a unique case study that can be used to apply the philosophical arguments involved in the free will debate to real agents whose illness may, at least partly, be the result of deterministic causal mechanisms. The agent has little to no control over most of these causal mechanisms, which seems to imply that she cannot have free will. However, I will attempt to argue that this is not necessarily the case. While the anorexic agent cannot control the causal mechanisms which contribute to her illness, she is able to retain control over her intentional actions. She possesses the capacity to reflect critically on her first-order preferences and desires and either identifies with these preferences or changes them via her higher-order desires and acts accordingly. As such, the power and choice are ultimately her own.

Keywords
Free will, determinism, compatibilism, anorexia nervosa

Developments in sciences such as biology, neuroscience, psychology, and psychiatry have been uncovering causal mechanisms which have improved our understanding of human character and behaviour, which seem to be the result of deterministic forces. This leads to questions regarding the status of free will as it is often regarded as incompatible with determinism. While the free will versus determinism debate has always been a favorite in philosophy, most of it is focused on the philosophical arguments themselves and not on how the philosophical positions actually apply in real-world cases. Many of the examples in the philosophical literature are thought experiments or fictional cases used to illustrate an argument, when it would be more useful to utilize a real-life example to see how these philosophical arguments pan out when they are applied. The free will debate gives us a unique opportunity to do just this, as psychological research supports philosophy's deterministic position by providing empirical evidence regarding human behaviour that bolsters its claims. One such area of research examines the causal mechanisms that contribute to anorexia nervosa. By using anorexic patients as a case study to examine the philosophical arguments which attempt to reconcile free will with determinism, we are able to examine agents who are practically useful and

philosophically interesting to study. Individuals with anorexia nervosa are influenced by real causal forces which present a philosophical challenge for one who wishes, as I do, to find a way to argue that despite these forces, the individuals with anorexia are still able to retain free will.

This paper will first present the philosophical position in support of a deterministic perspective and show how the empirical research on anorexia in psychology supports this philosophical claim. In the second section, philosophical attempts to reconcile free will with determinism (the compatibilist position) will be presented and shown why they are inadequate when applied to the real-life anorexic agent. Finally, the third section will explore a solution that considers the role autonomy plays in free will. It will be argued that the anorexic agent is indeed able to retain some free will despite the deterministic causes of her illness, for while she may not be able to control the causes which contribute to her illness, she is still able to retain power and control over her intentional actions. She is able to do this either by identifying with her predetermined preferences and permitting the behaviour or action to commence (for example, refusing to eat), or she may instead opt for another course of action that is of her choosing (such as deciding to eat). The power and choice is ultimately her own, and as such, her responsibility.

1. Why Determinism?

A Philosophical Perspective

Robert Kane gives a wonderful overview of the literature on the free will versus determinism debate in his introduction to the *Oxford Handbook of Free Will*, which I will draw from to set up my project. To begin the discussion one must note that modern debates on this topic stem from two questions: (1) ***the Determinism Question*** which asks "Is determinism true?" and (2) ***the Compatibility Question*** which wonders "Is free will compatible (or incompatible) with determinism?" (Kane 2009, 2). Kane explains that the "[a]nswers to these questions give rise to two major divisions in contemporary free will debates" between determinists and indeterminists for the first question, and compatibilists and incompatibilists for the second (2009, 2). Regarding the determinism question, the prevailing (though certainly not uncontested) view in modern quantum physics is that we live in an indeterministic world. Quantum theory "denies that elementary particles composing the "system of the world" have exact positions and momenta that could be simultaneously known by any such intelligence (the Heisenberg Uncertainty Principle); and it implies that much of the behavior of elementary particles ...

is not precisely predictable and can by explained only by probabilistic, not deterministic, laws" (Kane 2009, 2). "[T]he uncertainty and indeterminacy of the quantum world", continues Kane, "is not merely due to our limitations as knowers but to the nature of the physical world itself" (2009, 2).

However, while universal indeterminism is the predominate view in physics, deterministic views regarding human behaviour have been on the rise in other sciences such as biology, neuroscience, psychology, and psychiatry (Kane 2009, 3). There are many reasons for this trend, but the most relevant one for our present purpose is that empirical research in human behaviour and action has given us "greatly enhanced knowledge of the influence of genetics and heredity upon human behavior ... greater awareness of biochemical influences on the brain; the susceptibility of human moods and behaviour to drugs... influences of psychological, social, and cultural conditioning upon upbringing and subsequent behaviour, and so on" (Kane 2009, 3). These findings seem to be pointing more and more towards our behaviour being "determined by causes unknown [unconscious] to us and beyond our control" (Kane 2009, 3). As a result, as Sam Harris argues:

> Today, the only philosophically respectable way to endorse free will is to be a compatibilist – because we know that determinism, in every sense relevant to human behavior, is true. Unconscious neural events determine our thoughts and actions – and are themselves determined by prior causes of which we are subjectively unaware. However, the "free will" that compatibilists defend is not the free will that most people feel they have. (2012, 16)

We will get to the compatibilist position in the next section, but first we will look at what Harris means by 'free will that most people feel they have,' which is typically referred to as Libertarian Free Will.

This concept can be seen as the "freest" of the free will positions as it asserts free will exists and the universe is indeterministic. Contemporary free will libertarians must be able to successfully deny determinism as well as deny the compatibility of free will and determinism (Kane 2009, 8). Unlike the issue compatibilists face with reconciling free will with determinism (again, more on that shortly), the problem for the libertarian is how to make sense of a conception of free will that is *incompatible* with determinism by finding a way that we can have free will in an indeterministic world (Kane 2009, 8). The problem with indeterminism is that an event may or may not occur, it is a matter of chance and, as Kane explains, "chance events are not under the control of anything, hence not under the

control of the agent. How then could they be free and responsible actions?" (2009, 8). Instead of assisting the libertarian position by escaping the problems with determinism, indeterminacy may undermine freedom rather than enhance it (2009, 8).

Libertarianism was a worthwhile attempt, for as Saul Smilansky explains, "it was supposed to allow a deep moral connection between a given act and the person, and yet not fall into being merely an unfolding of the arbitrarily given, whether determined or random" (2009, 1-2). However, Smilansky concludes that this project is not plausible because, as Adina Roskies states "[t]his view does not seem to cohere with any scientific picture that we know" (2006, 420). This is the general philosophical consensus on libertarian free will, though there are certainly still philosophers working on the libertarian view, including Kane. Because psychological research supports a deterministic view of human behaviour and action, and libertarian free will goes against empirical studies on how we make our decisions, we will put aside the libertarian view and turn to the psychological research that supports the philosophical case for determinism.

A Psychological Perspective

While the exact etiology of anorexia nervosa is not known, the psychological research is uncovering some important findings that are making promising headway in uncovering the causal mechanisms that underlie this eating disorder. In this section I will present some examples from the research on the deterministic causal forces that contribute to anorexia. Before I do so the diagnostic criteria for anorexia nervosa will be outlined to give an overview of the sort of characteristics the illness exhibits.

Anorexia Nervosa: DSM 5 Diagnostic Criteria

1. Restriction of energy intake relative to requirements, leading to a significantly low body weight in the context of age, sex, developmental trajectory, and physical health. *Significantly low weight* is defined as a weight that is less than minimally normal or, for children and adolescents, less than that minimally expected.
2. Intense fear of gaining weight or of becoming fat, or persistent behavior that interferes with weight gain, even though at a significantly low weight.
3. Disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or persistent lack of recognition of the seriousness of low body weight.

Restricting Type

During the last 3 months, the individual has not engaged in recurrent episodes of binge eating or purging behavior (i.e., self-induced vomiting or the misuse of laxatives, diuretics, or enemas). This subtype describes presentations in which weight loss is accomplished primarily through dieting, fasting, and/or excessive exercise.

Research Findings

A well cited study by Goss and Gilbert (2002) suggests that shame may play a large role in the anorexic's eating disorder. The researchers assert the restrictive symptoms may function in a shame-pride cycle where "feelings of internal and external shame lead to restriction and the subsequent weight loss (successful restriction) leads to feelings of pride so a shame-pride cycle develops where shame negatively reinforces and pride positively reinforces the primary symptoms of restriction and weight loss" (Troop et al. 2008, 481). Shame is considered to be an emotion that has a highly social component for it is concerned with a fear or anticipation of eliciting disgust in others, either in the presence of a real or imagined audience (Troop et al. 2008, 480). It has strong implications with eating disorders and anorexia is more commonly associated with external shame, and bulimia tied to internal shame. External shame is the result of a person's perception that others view her in a negative manner, in that "the self is perceived by others as an object of scorn, ridicule and contempt" (Troop & Redshaw 2012, 373). Internal shame is a negative reflection of how one views oneself such as "worthless, flawed, morally defective or unattractive" (Troop & Redshaw 2012, 373). The difference between the two can be summarized as "the experience of being shamed versus feeling ashamed" (Gilbert 1998, as cited in Troop et al. 2008, 481). A proneness to experiencing shame is positively related to issues with eating in female students, and women who exhibit symptoms of an eating disorder experience more guilt and shame regarding eating than do depressed individuals and student controls (Troop et al. 2008, 481). While these findings were among non-clinical female samples where it was not established whether eating disorders were present or not, this research gives a good point of comparison as currently ill and recovered sufferers of eating disorders reported higher levels of shame (related to character and eating) than a student comparison group (Troop et al. 2008, 481). Additionally, the current sufferers of eating disorders report higher levels of shame regarding their character and eating than those who had recovered (Troop et al. 2008, 48). One study found that inpatients with anorexia and bulimia reported higher levels of

internalized shame than patients with depression and anxiety diagnoses (Grabhorn et al. 2006, as cited in Troop et al. 2008, 481).

Another causal factor that may contribute to anorexia nervosa is a body image dysfunction. There are two main types of body image dysfunctions that have been identified, a perceptual body-size distortion and cognitive-evaluative dissatisfaction. The former occurs "when a person has difficulty accurately gauging her body size", such as when the anorexic estimates her body size is larger than it actually is (Cash & Deagle 1997, 108). The latter is concerning one's attitudes about her body image and is often referred to as "body dissatisfaction" or "disparagement" (Cash & Deagle 1997, 108). In this case the anorexic may have an accurate view of her body size, but is quite dissatisfied with her body's size or shape. These two types of body image dysfunctions seem to be largely independent (Cash & Deagle 1997, 108). While women without an eating disorder can more or less accurately perceive their bodies, patients with anorexia overestimate their body fat (Benninghoven et al. 2007, 55). They also have been demonstrated to be less satisfied with their current body shape than the control subjects and have a thinner ideal body size in relation to what they see as their perceived body size (Mohr et al. 2009, 1524). Body satisfaction is calculated by asking subjects to select an image of a body type that matches their ideal body size, and another matching their actual body size. The discrepancy between the subjects' self-perception and their ideal perception is interpreted as a measure of body satisfaction (Mohr et al. 2009, 1521). When conducting a satisfaction rating with functional magnetic resonance imaging (fMRI) for thinner self-images, subjects with anorexia demonstrated stronger activation of the insula and the lateral anterior prefrontal cortex, while the controls "showed a stronger recruitment of precuneus during body size estimation for fatter images" (Mohr et al. 2009, 1524). Mohr et al. conclude from this that they "were able to separate the two different dimensions of body image behaviorally and find neural correlates with different task-specific involvements for anorectic patients and healthy controls" (2009, 1524). Furthermore, the researchers were able to find that anorexic patients had a higher activation of the left insula for the thin condition in the satisfaction rating in contrast to the thin condition in the body size estimation task (Mohr et al. 2009, 1525-1526). Results in brain areas and body image studies "indicate alterations in the activation of posterior parietal regions for patients with anorexia nervosa, possibly related to spatial components of the body image" (Mohr et al. 2009, 1520). Mohr et al. suggest that their findings could imply this insula activity could be associated with "a stronger experience of emotions by the patient group while viewing the self-images" (2009, 1527). Additionally, "[i]t could be speculated that the higher activity of the insula for thin self-images in the anorectic group is a

consequence of higher emotional valence of the thinner self-images" (Mohr et al. 2009, 1527) as insula activation has been tied to emotional and interoceptive awareness (Mohr et al. 2009, 1526).

Finally, the dysregulation of reward-processing mechanisms is believed to play a central role in eating disorders by contributing to and maintaining the core symptoms (Alonso-Alonso 2013, 1082). There appears to be a specific dysfunction with the suppression of the desire to eat ("wanting" food) while the capacity to evaluate ("liking") food is preserved (Berridge 2009, as cited in Alonso-Alonso 2013, 1082). A recent study by Frank et al. utilized voxel-based morphometry (VBM), a technique for automatic computational neuroanatomy, to examine the difference in brain structure in patients with or who had recently recovered from anorexia, patients with bulimia, and healthy control women (Alonso-Alonso 2013, 1082). The researchers found the anorexic and bulimic patients had larger gray matter volume in the gyrus rectus/medial orbitofrontal cortex, an area of the ventral system, compared to the controls (Alonso-Alonso 2013, 1083). The gray matter volume in this area correlated with pleasantness ratings ("liking") during a sucrose taste perception test prior to scanning (Alonso-Alonso 2013, 1083). The anorexic subjects "had increased gray matter volume in the anterior insula in the right hemisphere" while the bulimic subjects had a similar effect on the left side (Alonso-Alonso 2013, 1083). Furthermore, the recovered anorexic subjects and those with bulimia had reduced gray matter volume in the dorsal striatum, an area associated with a measure of reward sensitivity ("wanting") (Alonso-Alonso 2013, 1083). These research results find the gyrus rectus/medial orbitofrontal cortex as a neural substrate common to both anorexia and bulimia nervosa, and since this area is predominately involved in the processing of pleasant stimuli (Gabenhorst & Rolls 2011, as cited in Alonso-Alonso 2013, 1083), "this opens the possibility that brain changes determining an elevated capacity to experience pleasure from food (liking the taste) could represent neurodevelopmental contributors to eating disorders and act as an initial trigger of compensatory behaviors (such as decreased wanting for food rewards) early on in the natural history" (Alonso-Alonso 2013, 1083-1084).

These studies only present a very small sample of the research on anorexia nervosa but are intended to demonstrate the various mechanisms that may contribute to this eating disorder. Heightened emotional responses such as shame or a dysfunctional body image, or physiological issues such as a dysregulation in the reward centers of the brain are all possible deterministic factors over which the anorexic agent has no control. Her refusal to eat, and thus maintain a healthy body weight, is a much more complicated issue

than her simply declining to eat. This empirical data combined with the philosophical argumentation for determinism now directs us to the question of free will.

2. Pursuing Compatibilism

Having affirmed the Determinism Question[1] by confirming the causation of anorexia is indeed deterministic, I will now turn to the Compatibility Question [2] which is the true purpose of this paper – if the anorexic's illness is the result of deterministic processes, can she somehow maintain her free will? Sam Harris argues she cannot, for "[h]ow can we be "free" as conscious agents if everything that we consciously intend is caused by events in our brain that we *do not* intend and of which we are entirely unaware?" (2012, 25-26). Free will can be seen as incompatible with determinism because "[t]o say that "my brain" decided to think or act in a particular way, whether consciously or not, and that this is the basis for my freedom, is to ignore the very source of our belief in free will: the feeling of *conscious* agency. People *feel* that they are the authors of their thoughts and actions, and this is the only reason why there seems to be a problem of free will worth talking about" (Harris 2012, 26). However, the Compatibility Question does not *assume* that if determinism is true then we necessarily lack free will, and the burden of proof lies with those who assert the two are incompatible (Kane 2009, 3). The question does, however, seem to *imply* incompatibility (and historically this was the assumption) for when we speak of the ability to make a decision we assume we are able to choose from a variety of different possibilities (Kane 2009, 4). Modern arguments for incompatibilism often stem from this assumption, "the requirement that an agent acted freely, or of his or her own free will, only if the agent had *alternative possibilities*, or *could have done otherwise*" (Kane 2009, 4). Kane (2009, 4) refers to this as the **Alternative Possibilities (AP) Condition** which is as follows:

> The case for incompatibility from this Alternative Possibilities Condition has the following premises:
>
> 1. The existence of alternative possibilities (or the agent's power to do otherwise) is a necessary condition for acting freely (of one's free will).

1. Is determinism true?

2. Is free will compatible (or incompatible) with determinism?

2. Determinism is not compatible with alternative possibilities (it precludes the power to do otherwise).

A modern, and widely discussed, argument in support of premise 2 is the Consequence Argument, which is informally stated by van Inwagen (1983) as follows:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (as cited in Kane 2009, 4)

"Up to us" is the wording that Kane also favors, for we feel that it is "up to us" "what we choose and how we act" (Kane 2009, 1). Just as we could not choose events in the past we also cannot choose the impact these past actions will have on our present and future acts. It thus appears we cannot do anything to alter our present actions as we do not have the power to do otherwise, and hence lack alternative possibilities (Kane 2009, 4). In order to hold a compatibilist position one must successfully defeat the Consequences Argument, in addition to other incompatibilist arguments. We will now turn to the compatibilist position.

Classical compatibilists often define freedom in terms of one having the power or ability to act. It then follows that in order to be free one must have (1) the power and ability to act as one chooses or desires which requires (2) an "absence of constraints or impediments" which prevent one from doing as he wills, desires, or chooses to do (Kane 2009, 4). Traditional examples of constraints or impediments include things like "physical restraints, lack of opportunity, duress or coercion, physical and mental impairment" and so on (Kane 2009, 4). Kane notes "[y]ou lack the freedom to meet a friend in a cafe across town if you are tied to a chair, are in a jail cell, lack transportation, someone is holding a gun to your head, or you are paralyzed" (2009, 4-5).

However, for our purposes here we are more interested in subtle, biological constraints or impediments such as dysfunctional reward processing mechanisms that prevent the anorexic from making a truly free choice. In order for an agent to have free will that is compatible with determinism the agent must still be able to choose despite deterministic conditions. For our anorexia case, what would such a choice look like? Benjamin Libet has an interesting reply that focuses on the role that consciousness plays in voluntary action. Libet conducted an experiment where the subjects flicked their wrists "at any time they felt the urge or wish to do so" and were able to perform this flicking

of the wrist free of any external limitations or restrictions (for example, without time limitations or specific timed intervals) (Libet 2009, 1). He discovered that the readiness potential, which is "a specific electrical charge in the brain," began 550 milliseconds before the flicking of the wrist (2009, 1). The subjects were not aware of the intention to act until 350-400 milliseconds – which is *after* the readiness potential starts, but their conscious awareness occurs 220 milliseconds *before* the motor action (Libet 2009, 1). So while the volitional process is initiated unconsciously, the conscious function can still, according to Libet, control the outcome (whether to flick one's wrist or not) (Libet 2009, 1). Libet concludes that this means free will is not excluded as conscious will appears around 150 milliseconds before the muscle is activated (even though it does follow the unconscious onset of the readiness potential) (2009, 2). This unconscious onset is not problematic for Libet, as we still have enough time in the 150 milliseconds to either allow or veto the muscle action. Libet states he was "able to show that subjects could veto an act planned for performance at a prearranged time. They were able to exert the veto within the interval of 100 to 200 msec, before the preset time to act" (Libet et al., 1983b, as cited in Libet 2009, 3). He continues that "[a] large RP [readiness potential] preceded the veto, signifying that the subject was indeed *preparing* to act, even though the action was aborted by the subject" (2009, 3). Libet argues that "the conscious veto may *not* require or be the direct result of preceding unconscious processes. The conscious veto is a *control* function, different from simply becoming aware of the wish to act" (2009, 4).

Philosophically, this certainly seems like a promising compatibilist position. The veto power successfully defeats incompatibilism via the Alternative Possibilities Condition by defeating the second premise:

1. The existence of alternative possibilities (or the agent's power to do otherwise) is a necessary condition for acting freely (of one's free will).
2. ~~Determinism is not compatible with alternative possibilities (it precludes the power to do otherwise).~~

However, psychologically the veto is still problematic as we are left having to explain the anorexic who is able to veto her desire to refuse food and can eat (thus can overcome her illness) versus the one who decides not to stop the desire to refuse food and isn't able to gain weight (and in turn, get better). Why is one successful with the veto power and the other is not? Does the anorexic who ultimately refuses her veto power want to encourage her illness, or does she just have a weaker veto (willpower) than the anorexic who is able to get better? Either way, Libet's account leaves these questions unanswered

and is thus unsatisfactory for our purposes. If we are to grant the anorexic agent some control over her illness we cannot do it via Libet's account and need to look elsewhere.

Smilansky presents a more radical solution to the problem of free will in a deterministic world. He finds libertarian free will impossible but also finds too many problems with compatibilism to go that route. He argues "if there is no libertarian free will, no one can be ultimately in control, ultimately responsible, for this self and its determinations…[i]f people lack libertarian free will, their identity and actions flow from circumstances beyond their control" (2009, 2). It is due to this lack of control that the compatibilist has to accept "a shallower sort of meaning and justification" for moral worth, and it is due to this shallowness and "complacent compliance with the injustice of not acknowledging lack of fairness and desert" that Smilansky disagrees with compatibilism (2009, 2).

While Smilansky is more in favor of hard determinism which denies free will, he does not fully endorse it either. The hard determinist would assert that responsibility is a non-issue for we deserve neither praise nor blame for our actions, as ultimately they are beyond our control. However, Smilansky argues that rejecting notions of responsibility makes the hard determinist "morally blind and a danger to the conditions for a civilized, sensitive moral environment" (2009, 4). Instead what Smilansky wants is a Community of Responsibility that is comprised of members whose choices determine the moral attitudes they receive (2009, 3). Smilansky argues that the assumption that one must either affirm compatibilism or incompatibilism is incorrect, and states that while the two are logically inconsistent this does not mean one cannot hold an intermediate and mixed position between the two (2009, 2). By rejecting the assumption that only compatibilism or incompatibilism can be correct we are able to stay closer "to the deepest issues on the free will issue" and "proceed along a new path that ultimately runs closer to the intuitive field than do either of the conventional monisms" (Smilansky 2009, 2). The rejection of the Assumption of Monism paves the way for the Community of Responsibility, which is achieved via Illusionism. This is "the position that illusion often has a *large and positive role* to play in the issue of free will" for we are "fortunately" deceived into thinking we have free will and this allows us to maintain "civilized morality and personal value" (Smilansky 2009, 5).

Smilansky is not suggesting we induce illusionary beliefs regarding free will, or maintain ones that we "fully realize" are illusionary (2009, 4). Instead we are to permit and humour the illusionary beliefs that are already in place, for they are doing more good than harm. "The sense of "illusion" that I am using", explains Smilansky, "combines the falsity of a belief with some motivated role in forming and maintaining that belief

– as in standard cases of wishful thinking or self-deception" (2009, 5). Smilansky fully acknowledges that these beliefs are false, for "[a]ll our actions, however an internalized and complex a form they make take, are the result of what we are, ultimately beyond our control" (2009, 6). However, because compatibilism and hard determinism, for Smilansky, are too problematic to endorse on their own, illusionism is the only solution to maintain some sense of moral order and responsibility in this deterministic world.

Smilansky's illusionism is bound to be highly criticized, and for good reason. His motivation for a Community of Responsibility is understandable as it aligns with the moral intuition that agents deserve praise or blame for their actions. However, while we all surely entertain false but useful beliefs in our everyday lives – "Our troubles will soon end, for tomorrow will be better than today" etc. – actively encouraging them as a philosophical position is too problematic to endorse. The problem is illusionism is essentially a cop out, a last-ditch effort to salvage free will if only in name. Is the prospect of denying free will so devastating to us as agents that we must retain some semblance of free will, even in name only? Sam Harris doesn't think so; in fact he asserts that instead of becoming fatalistic as a result of denying free will it has actually increased his feeling of freedom. He states "[m]y hopes, fears, and neuroses seem less personal and indelible. There is no telling how much I may change in the future" (Harris 2012, 46). One need not draw lasting conclusions about himself based on how he thought or behaved in the past, for "[a] creative change of inputs to the system – learning new skills, forming new relationships, adopting new habits of attention – may radically transform one's life" (Harris 2012, 46). This all sounds quite promising, but we have to remember that if we deny free will these changes of inputs to our system would have to be predetermined as well and I am not ready to throw in the towel just yet. We will look at one more account before we call it quits for the day.

Alfred Mele's discussion of autonomy and self-control may help provide additional insight to the task at hand, for he uses autonomy as an extension of free will. In order for one to be autonomous they must be able to choose or act freely, to some degree. Furthermore, they must have some level of self-control, which Mele states agents possess when they "have significant motivation to conduct themselves as they judge best and a robust capacity to do what it takes so to conduct themselves in the face of (actual or anticipated) motivation" (Mele 1995, as cited in Mele 2009, 2). However, even if one is an "ideally self-controlled" person and is thus able to manifest "perfect self-control"[3] Mele

3. Mele states such a person would achieve the four dimensions for perfect self-control: range, object, frequency, and effectiveness. For more detail see Mele 1995.

argues this still insufficient for autonomy. Mele uses Gerald Dworkin's explanation of autonomy to explain why, which is as follows:

> [A]utonomy is a second-order capacity to reflect critically upon one's first-order preferences and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values. (Dworkin 1988, as cited in Mele 2009, 3)

Mele argues "[a]n ideally self-controlled person has this capacity and ability. However, even ideal self-control – no matter how frequently and successfully exercised – might not suffice for autonomy. If, as it seems, every process of critical reflection is regulated or guided by principles or values already in place, some principle or value will be presupposed or taken for granted in each process" (2009, 3). However, he does not think that this necessitates that we are unable to have any control if the world is deterministic. To assert "the thesis that there is at any instant exactly one physically possible future"[4] does not mean that there cannot be "*more than one physically possible future*" as causation does not depend, argues Mele, on the absence of these physically possible futures (2009, 3, my italics). Just because there is only one outcome does not necessitate that there were not any other possibilities that could have also been potential candidates for the final outcome.

Things get even more interesting when one examines the impact an internalist versus an externalist view of autonomy has on an agent's history. The internalist view only sees an agent's history as relevant to his autonomy in so long as it "yield[s] rationality, an ability to acquaint oneself with relevant facts, reliable capacities for decision-making and action, current psychic integration, and the like" (Mele 2009, 4). Mele explains "[g]iven that the traits and capacities are in place and are exercised with appropriate care and suitable frequency, all else is irrelevant to psychological autonomy, including how the agents came to be as they are" (2009, 4). Dworkin continues his view on autonomy mentioned above - "[A]utonomy is a second-order capacity to reflect critically upon one's first-order preferences and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values" (Dworkin 1988, as cited in Mele 2009, 3) by explaining that "[b]y exercising such a capacity we define our nature, give meaning and coherence to our lives, *and take responsibility for the kind of person we are*" (Dworkin 1988, as cited in Mele 2009, 5, my italics). Internalism views the cause of one's capacities and abilities as an independent issue from whether one performs an

4. This is Peter van Inwagen's (1983) definition of determinism, as cited in Mele 2009, 3.

act freely and thus is morally responsible for the action (Mele 2009, 5). This separation of cause and effect could be promising, as it would allow the compatibilist to assert that while the initial or background causes (one's history) are deterministic, free will lies in the intermediate step between the cause and effect. This is similar to Libet's veto power but with a longer interval than a few milliseconds to make changes. This intermediate step is when one critically reflects on one's first-order preferences and desires and decides whether to act in accordance with these desires, or to pursue a different course of action. Acting in accordance with the desire or pursuing a different action would be the effect, and it does not directly flow from the deterministic causal forces. It is one's second-order choice, and thus, he is responsible for his choice as he made it autonomously.

However, internalism is not without its critics, for the externalist view of autonomy is very interested in how an agent came to be. In this view autonomy depends on the agent's causal history, such as how she came to possess the desires and values that guide her self-reflection and decision making (Mele 2009, 4). The concern here becomes evident when two agents possess all of the (non-historical) qualities the internalist would require – reviewing facts, using reason in the decision-making process, critically examining their first-ordered desires, etc. – but their histories are so radically different that "we would be strongly inclined to regard one as significantly less autonomous than the other" (Mele 2009, 5). For example, say Patient A has a high level of external shame and frequently gets caught in a shame-pride cycle. The shame she feels from what she perceives as other peoples' negative perceptions of her leads her to restrict her food intake, which results in significant weight loss. This weight loss encourages her to continue restricting her food as she feels proud of her reduced body weight and control over her body. Patient B has difficulty accurately gauging her body size due to an assessment distortion in her self-perception, in addition to experiencing a decreased desire for food as a result of a dysregulation in the reward-processing mechanisms in her brain. Differences in histories such as these could help explain why some anorexics are able to get better and others are not, because some anorexic's histories (the deterministic events that lead to her illness) are too much for her to overcome – the deterministic deck is stacked against her, so to speak. When the factors that comprise one's causal history make free will seem implausible, is it possible to retain an internalist notion of free will that is more than just wishful thinking?

3. A Solution

Dworkin and Mele argue that in order to have autonomy we do not have to be completely independent of deterministic forces and I agree. There may be a way to reconcile the internalist and externalist views of autonomy. We can argue, like Mele, that if determinism is true and it is compatible with personal autonomy and free will we may not be able to take responsibility for our character but one can still have autonomy by "living in accordance with preferences and desires that one identifies with "in light of higher order preferences and values"" (Mele 2009, 6). We do not assume responsibility for our character as it is the product of external and deterministic causes over which we have no control and thus, no responsibility (Mele 2009, 6). However, these deterministic causes, our histories, need not undermine compatibilism for we still retain power (and responsibility) for our intentional actions. We may not choose what we've got, but we do choose what we do with it.

How does one exert control over her illness? A major step would be to seek treatment. Anorexia continues to have a poor prognosis (Galsworthy-Francis & Allan 2014, 55) and as such, the statistics for anorexics are grim. Steinhausen (2002) reports "[l]ongitudinal research has suggested fewer than 50% of individuals diagnosed with AN [anorexia nervosa] recover fully; 20–30% continue to experience residual symptoms, 10–20% remain significantly ill and 5–10% die from their illness" (as cited in Galsworthy-Francis & Allan 2014, 55). Morris (2008) states the mortality rates in anorexia are ten times that of the general population, and these are "the highest of all psychiatric disorders" (Harris & Barraclough 1998, as cited in Galsworthy-Francis & Allan 2014, 55). These statistics could be viewed as an agent's lack of free will or control over the illness; however it is interesting to note that individuals with anorexia tend to resist treatment and participation in treatment studies (Agras 2010, 488). Those who do seek treatment have a high rate of premature termination, the literature documents that rates of 50% are not uncommon (Sly et al. 2014, 40). It is beyond the scope of this paper to explore why premature termination rates are so high, as what is relevant and important for the task at hand is to demonstrate the anorexic *is able to make a choice*. If she can choose, opt, or decide whether or not to participate in treatment for her illness then she has free will as she has the ability to choose otherwise or contra to her first-order desires. The patients that prematurely terminate treatment typically do so at lower weights than those who complete treatment, "which is an indicator for the need for subsequent rapid readmission to [a] hospital" (Sly et al. 2014, 40). What is important to note is that these

patients *chose* to leave or not comply with treatment.[5] This choice defeats the second condition in the second premise of the Alternative Possibilities Condition:

1. The existence of alternative possibilities (or the agent's power to do otherwise) is a necessary condition for acting freely (of one's free will).
2. ~~Determinism is not compatible with alternative possibilities (it precludes the power to do otherwise).~~

About half of anorexic patients do get better or at least complete treatment – and half do not. The deterministic cause of one's illness does not preclude her from choosing from alternative possibilities - either to get better, or worse, or stay the same. This brings us back to Mele's assertion about physically possible futures. While only one of these outcomes is logically possible, there is nothing that necessitates which of these three outcomes will be determined because the choice is ultimately the anorexic's. She possesses the capacity to reflect critically on her first-order preferences and desires (such as to refuse food) and either identify with these desires or to change them (eat enough food so she is able to gain weight) via her higher-order desires. This is Dworkin and Mele's definition of autonomy and the anorexic fits the criteria.

This may actually be closer to Harris' argument than one would have initially thought. He argues:

> Becoming sensitive to the background causes of one's thoughts and feelings can – paradoxically – allow for greater creative control over one's life. It is one thing to bicker with your wife because you are in a bad mood; it is another to realize that your mood and behavior have been caused by low blood sugar. This understanding reveals you to be a biochemical puppet, of course, but it also allows you to grab hold of one of your strings: A bite of food may be all that your personality requires. Getting behind our conscious thoughts and feelings can allow us to steer a more intelligent course through our lives (while knowing, of course, that we are ultimately being steered). (2012, 47)

5. In some cases the premature termination is a staff initiated discharge which happens in cases where "the clinical team feel the patient is not engaging with the ethos of treatment, or not working in alliance with the boundaries of the treatment program. For example, a patient may disengage from the program and deliberately cease gaining weight" (Sly et al. 2014, 40). While the termination may be the decision of the clinical team, the choices that lead to this decision are the patient's.

This is along the lines of reconciling internalism and externalism, with the exception of one main difference – responsibility. I am willing to grant that the power of being able to "grab one of your strings" also gives you the ability to warrant responsibility for your actions. As such, we are able to remain in Smilansky's Community of Responsibility after all, but with a better justification than merely entertaining false beliefs. This, in turn, also permits the anorexic responsibility over her deterministic illness. She certainly did not choose or have any power over the underlying mechanisms that caused her eating disorder, but because she can choose her intentional actions she can be responsible for them. This of course seems more agreeable (and less problematic) when she is able to complete treatment and overcome her illness than if she is unsuccessful. But if we are able to attribute merit to an agent when she is successful, we must also attribute blame where it is warranted. The anorexic's autonomy and control can allow her to beat her illness. And indeed, some do.

References

Agras, Walter Stewart. 2010. "Chapter 27: Overview." In *The Oxford Handbook of Eating Disorders*, edited by W. Stewart Agras, 486–490. New York: Oxford University Press.

Alonso-Alonso, Miguel. 2013. "Brain, Reward, and Eating Disorders: A Matter of Taste?" *American Journal of Psychiatry* 170 (10): 1082–1085.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders* Fifth Edition. Washington: American Psychiatric Publishing.

Benninghoven, Dieter, Lena Raykowski, Svenja Solzbacher, Sebastian Kunzendorf and Gunter Jantschek, 2007. "Body Images of Patients with Anorexia Nervosa, Bulimia Nervosa and Female Control Subjects: A Comparison with Male Ideals of Female Attractiveness." *Body Image* 4 (1): 51–59.

Cash, Thomas F. and Edwin A. Deagle, 1997. "The Nature and Extent of Body-Image Disturbances in Anorexia Nervosa and Bulimia Nervosa: A Meta Analysis." *International Journal of Eating Disorders* 22 (2): 107–125.

Galsworthy-Francis, Lisa and Steven Allan. 2014. "Cognitive Behavioural Therapy for Anorexia Nervosa: A Systematic Review." *Clinical Psychology Review* 34 (1): 54–72.

Harris, Sam. 2012. *Free Will*. New York: Free Press.

Kane, Robert. 2009. "Introduction: The Contours of Contemporary Free Will Debates." *The Oxford Handbook of Free Will Online*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195178548.001.0001/oxfordhb-9780195178548-e-1.

Libet, Benjamin. 2009. "Do We Have Free Will?" *The Oxford Handbook of Free Will Online*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195178548.001.0001/oxfordhb-9780195178548-e-25.

Mele, Alfred R. 2009. "Autonomy, Self-Control, and Weakness of Will." *The Oxford Handbook of Free Will Online*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195178548.001.0001/oxfordhb-9780195178548-e-24.

Morh, H.M., J. Zimmermann, C. Röder, C. Lenz, G. Overbeck and R. Grabhorn. 2009. "Separating Two Components of Body Image in Anorexia Nervosa Using fMRI." *Psychological Medicine* 40 (9): 1519–1529.

Roskies, Adina. 2006. "Neuroscientific Challenges to Free Will and Responsibility." *Trends in Cognitive Sciences* 10 (9): 419–423.

Sly, Richard, Victoria A. Mountford, John F. Morgan and J. Hubert Lacey. 201). "Premature Termination of Treatment for Anorexia Nervosa: Differences Between Patient-Initiated and Staff-Initiated Discharge." *International Journal of Eating Disorders* 47 (1): 40–46.

Smilansky, Saul. 2009. "Free Will, Fundamental Dualism, and the Centrality of Illusion." *The Oxford Handbook of Free Will Online*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195178548.001.0001/oxfordhb-9780195178548-e-22.

Troop, Nicholas A and Chloe Redshaw. 2012. "General Shame and Bodily Shame in Eating Disorders: A 2.5-Year Longitudinal Study." *European Eating Disorders Review* 20 (5): 373–378.

Troop, Nicholas A., Steven Allan, Lucy Serpell and Janet L. Treasure. 2008. "Shame in Women with a History of Eating Disorders." *European Eating Disorders Review* 16 (6): 480–488.

Journal of Cognition and Neuroethics

Simply Irresistible: Addiction, Responsibility, and Irresistible Desires

Marcela Herdova
Florida State University

Biography
Marcela Herdova is Postdoctoral Research Fellow in Self-Control at Florida State University. She previously worked as Research Associate on the "Self-Control and the Person: A Multi-Disciplinary Account" project at King's College London where she also earned her PhD in 2011. Her research interests are action theory, free will, moral psychology, consciousness and applied ethics.

Acknowledgements
I would like to thank Stephen Kearns for his very helpful comments on the earlier drafts of this paper. I would also like to thank the audience at the CUNY Cognitive Science Speaker Series as well as the participants at the Free Will conference at the University of Michigan-Flint.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Herdova, Marcela. 2015. "Simply Irresistible: Addiction, Responsibility, and Irresistible Desires." *Journal of Cognition and Neuroethics* 3 (1): 195–216.

Simply Irresistible: Addiction, Responsibility, and Irresistible Desires

Marcela Herdova

Abstract

In this paper I set out to investigate the claim that addicts lack sufficient control over their drug-taking and are thus not morally responsible for it. More specifically, I evaluate what I call the Simply Irresistible Argument, which proceeds from the claim that addictive desires are irresistible to the conclusion that addicts are not responsible for acting on such desires. I first propose that we have to disambiguate the notion of an irresistible desire according to temporal criteria, and revise the original argument accordingly in two different ways; one involving proximally irresistible desires and one involving permanently irresistible desires. I propose that both versions of the Simply Irresistible Argument fail, and, as a result, that considerations about irresistible desires and control cannot extricate addicts from responsibility for their drug-taking.

Keywords

Addiction, control, irresistible desires, moral responsibility

1. Introduction

Debates about drug addiction mainly center around three interrelated issues: what addiction is, how to treat addiction, and the moral and legal responsibility of addicted individuals. I shall focus here on the issue of moral responsibility—namely on an argument that addicts lack sufficient control over their drug-taking and are thus not morally responsible for it.

The argument I evaluate, which I dub the *Simply Irresistible Argument*, builds on the assumption that addictive desires are *irresistible*. After introducing this argument, I disambiguate the notion of an irresistible desire according to temporal criteria and reconstruct the original argument in two different ways (in light of this disambiguation). I propose that both versions of the argument fail, and that considerations about irresistible desires and control cannot extricate addicts from responsibility for their drug-taking. However, to conclude, I also make a distinction between control (and responsibility) for individual actions, on the one hand, and for general long-term patterns of behavior, on the other. I propose that an agent's controlling and being responsible for individual actions does not entail her controlling and being responsible for the pattern of behavior made up of those actions. As a result, it might be the case that an addict's

individual drug-taking actions and her general condition (addiction) deserve, from a moral perspective, rather different treatments.

Before I present the Simply Irresistible Argument, I shall set out some of the assumptions I am making about control and responsibility, and discuss the importance of irresistible desires to control.

1.1 Preliminary Remarks: Control and Responsibility

There are two broad approaches to moral responsibility: volitionist and non-volitionist. On a volitionist approach, moral responsibility requires that the agent is in *control* of her behavior. An agent has diminished or even no responsibility for behavior which she does not control in the right manner. A non-volitionist, on the other hand, does not impose a control condition on moral responsibility—one might be held accountable even for those actions which one cannot control or choose. Instead, non-volitionists propose that phenomena other than control ground responsibility. For instance, Angela Smith proposes, on her rational relations view, that "To say that an agent is morally responsible for something … is to say that that thing reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it" (Smith 2008, 369).

In this paper, I shall assume that something like a volitionist approach to moral responsibility is correct. The Simply Irresistible Argument for the exculpation of addicts, as well as my response to it, are both based on such an approach.[1] I will also assume, as is widely accepted, that both control and responsibility come in degrees (e.g. Sinnott-Armstrong 2013). In other words, neither control nor responsibility are absolute notions—they come on a spectrum and one can have more or less of each. Relatedly, I shall also assume that degrees of control map onto degrees of responsibility. If one's control over one's behavior is diminished, so will be one's responsibility for this behavior. In general, actions over which an agent exercises greater degrees of control will warrant our attributing greater degrees of responsibility to her, and less control warrants attributing less responsibility.[2]

1. It should be noted, however, that a non-volitionist can construct her own versions of the Simply Irresistible Argument according to which irresistible desires provide the basis for exculpating addicted individuals on grounds other than diminished control. While the traditional debate on addiction, control and responsibility is couched in volitionist terms, such non-volitionist arguments certainly deserve further consideration. It is beyond the scope of my paper to discuss or evaluate these arguments here.

2. For present purposes, I shall not engage in the debate about the significance of determinism to control

1.2 Types of Control

How are irresistible desires connected to control (and thus, given our volitionist assumption, to responsibility)? There are different, related, types of control. These include, for instance, (i) reasons-responsiveness (the ability to recognize and react to reasons for action); (ii) the ability to do otherwise (the agent is able to, minimally, refrain from acting as she did); (iii) the translation of long-term commitments and values into action; (iv) authorship (the agent's actions are appropriately connected to her character); and (iv) self-control (the ability to control wayward motivations). Impairment of any of these types of control might, in part, explain (excessive) drug-taking behavior. For instance, a failure to recognize the danger of addictive substances, or a diminished ability to resist addictive desires, or a flat-out inability to refrain from acting on such desires, can help explain why addicts start (and continue) to use drugs despite undesirable consequences.

What might impair and diminish the above types of control? Some common such factors (in relation to addiction) include various cognitive biases which influence what reasons we attend to or ignore, and how we weigh and weight reasons. Automatization of behavior (while often increasing one's control) may also decrease various types of control, since it might result in the bypassing of an agent's rational and deliberative capacities in an undesirable way. Additionally, an overall reduction in know-how, skills and general mental abilities will often result in impaired control due to the fact that an agent may not recognize or be able to utilize various (self) control methods.

One factor which may have an especially wide-ranging and strong influence, diminishing or even completely eradicating various types of control, are desires with great motivational strength.[3] Such desires may affect one's reasons-responsiveness (by, for example, affecting one's willingness to consider certain reasons, or straightforwardly undercutting the motivational strength of other relevant reasons/desires), or authorship of one's actions (if the strong desires that move one to action are out of character), or

---

and responsibility. Determinism is the view that the past, together with laws of nature, necessitate any future events; i.e. there is one possible future given the laws of nature and the past. Some think that if determinism is true, no one is responsible for anything. My interest is less broad—I am concerned with whether addiction is a special threat to responsibility, assuming that non-addicts are generally responsible for their behavior.

3. Motivational strength, i.e. the strength with which desires move one to action, should not be confused with affective strength. Affective strength refers to the "felt" or "experienced" strength of desires. It is not obvious if, or to what degree, motivational strength correlates with affective strength (for a discussion about this, see Mele 2014).

even one's ability to do otherwise, rendering the agent unable to act differently than she in fact did.

If an agent has an *irresistible* desire, that is, a desire with such great motivational strength that it *compels* her to action, at least one type of the agent's control is diminished to the greatest degree—her ability to do otherwise. By definition, irresistible desires are those that one cannot refrain from acting on; they undercut any method of control that one may successfully utilize against resistible desires.

One seemingly plausible argument, then, that addicts are not responsible for their drug-taking appeals to their having irresistible desires (to take drugs). I introduce such an argument in section 2, and propose, in section 3, that we have to disambiguate this argument. Doing so provides us with two versions of this argument. In sections 4, 5 and 6 I explain why both arguments fail. In section 7 I consider the differences between control over individual actions and patterns of behavior, and then conclude in section 8 by considering some further alterations to the main arguments.

2. The Simply Irresistible Argument

Consider the following argument (in which "addicts" is short for "drug addicts"):

1. Addicts have irresistible desires to take drugs.
2. If addicts have irresistible desires to take drugs, then addicts are not (morally) responsible for taking drugs (when they do so as a result of such irresistible desires).
3. So addicts are not responsible for taking drugs.

We might understand premises 1 and 2 to be talking about *all* drug addicts or just *some* (though still, presumably, a significant number). For charity's sake I shall take "addicts" to refer to simply some significant number of addicts. Premise 1 is *prime facie* plausible. On some models of drug addiction, addiction is a disease (e.g. Charland 2002). Sufferers of this disease are compelled to act on pathologically strong desires for the drug in question. If addicts are truly compelled by such desires, then these desires are irresistible—the addict literally cannot resist acting on them. Premise 1 simply claims that this is true of some addicts, even if not all.

Premise 2 is also plausible on its face. If addicts really do have irresistible desires to take drugs, then they cannot do otherwise than take drugs. One can appeal to a popular and intuitive condition on moral responsibility; the Principle of Alternative Possibilities:

> (PAP) An agent is morally responsible for an action only if she could have done otherwise.

It is, it seems, unfair to hold someone responsible for something they could not help doing. Thus if a drug addict could not help but take drugs, she should not be held responsible for doing so. After all, our practices of moral responsibility are, typically, somewhat sensitive to similar considerations in cases other than addiction. Take, for instance, reflex behavior or various bodily tics. If someone spills water as a result of a bodily spasm or shouts an obscenity which is a manifestation of her tic, it usually weighs in on our assessment that this agent was unable to refrain from behaving the way she did. Such cases warrant different reactions than those in which agents could have refrained from such behavior (for instance, if one spills water on purpose, or intentionally insults someone by swearing at them). An agent is off the hook if she *couldn't help it*. PAP, or variations of it, reflect the importance many of us take this type of control to have for responsibility.

Since Frankfurt 1969, PAP has been vigorously attacked and equally staunchly defended. Frankfurt imagined scenarios such as the following (though this example is Fischer's):

> [Black] has secretly inserted a chip in Jones's brain that enables Black to monitor and control Jones's activities. Black can exercise this control through a sophisticated computer that he has programmed so that, among other things, it monitors Jones's voting behavior. If Jones were to show any inclination to vote for McCain (or, let us say, anyone other than Obama), then the computer, through the chip in Jones's brain, would intervene to assure that he actually decides to vote for Obama and does so vote. But if Jones decides on his own to vote for Obama (as Black, the old progressive would prefer), the computer does nothing but continue to monitor—without affecting—the goings-on in Jones's head.
>
> Now suppose that Jones decides to vote for Obama on his own, just as he would have if Black had not inserted the chip in his head. It seems, upon first thinking about this case, that Jones can be held morally responsible for his choice and act of voting for Obama, although he could not have chosen otherwise and he could not have done otherwise. (Fischer 2010, 316)

In essence, Jones could not have done otherwise than he did because Black's computer is waiting in the wings to make him (decide to) vote for Obama should Jones show any inclination not to. Since, in the end, Jones shows no such inclination and votes for Obama on his own, Jones is morally responsible for doing so, even though he cannot do otherwise. PAP is thus false.

Whether or not this counterexample works against PAP, it is compelling enough to damage the above argument for premise 2. If PAP has plausible counterexamples, then an agent may well be responsible for an action despite being unable to do otherwise. Still, Frankfurt goes on to explain that what does the work in his counterexamples to PAP (assuming they are successful) is that the element that renders the agents unable to do otherwise plays no role in *causing* the agent to act. Frankfurt suggests the following alternative to PAP:

> (PAP2) An agent is not morally responsible for an action if she performs the action *only because* she cannot do otherwise. (compare Frankfurt 1969, 838)

The Jones case does not falsify PAP2. While the computer renders Jones unable to do otherwise, it is not true that Jones votes for Obama *because* he cannot do otherwise.

When an agent acts on an irresistible desire, on the other hand, the desire itself compels her to action. She acts on this desire precisely *because* she cannot do otherwise. Indeed, she acts on this desire *only* because she cannot do otherwise (that is, she would act on this desire whether she wanted to or not—her having this desire, and her acting on it, are quite insensitive to the agent's wishes). Given this, addicts who take drugs as a result of irresistible desires to do so are not morally responsible for such actions. This is a plausible result that, further, does not rely on the controversial version of the principle of alternative possibilities, PAP, but rather than on the much more plausible PAP2.

3. Distinguishing Irresistible Desires

The Simply Irresistible Argument has, then, much to be said for it. Despite this, I believe that it fails. To see why, we need to discern an ambiguity in the argument concerning the idea of an irresistible desire. Desires, resistible or irresistible, can be about immediate or near immediate courses of action, or about future courses of action. Now consider a further, and crucial, distinction we can make concerning irresistible desires of

the former kind (desires to act immediately or very soon), based on the *time frame in which such desires are irresistible*:[4]

> **Proximally Irresistible Desires**: desires (to act immediately or very soon) which one cannot, after they have arisen, suppress or prevent oneself from acting on.
>
> **Distally Irresistible Desires**: desires (to act immediately or very soon) which one could not beforehand prevent from arising *and* could not beforehand prevent oneself from acting on.
>
> **Permanently (proximally and distally) Irresistible Desires**: desires (to act now or very soon) which one could not beforehand prevent from arising and which one could not beforehand, and cannot after they have arisen, suppress or prevent oneself from acting on.

The idea of a proximally irresistible desire is, I take it, the one most familiar to us. An agent has a desire to act now (or soon); she cannot rid herself of such a desire, and she cannot prevent herself from acting on it (either by intentionally resisting temptation or by simply doing something else instead). Common, but controversial, examples of such desires are those of people with OCD to perform various tasks (such as to wash their hands), and those of kleptomaniacs to steal. On some pictures of action, all such agents are compelled to act on their desires, and cannot stop themselves from doing so once such desires arise. (Any such examples will remain controversial, however, given the lack of compelling empirical evidence in support of such irresistible desires. Notwithstanding the empirical evidence, an agent with such proximally irresistible desires is easily conceivable).

Distally irresistible desires are somewhat less talked about. Such a desire is one that the agent cannot beforehand prevent from arising or from leading to action once it does arise. It is simple enough to think of cases in which an agent is unable to prevent a desire from *arising*. Though Bob is not hungry now, he will be. He cannot now prevent his desire to eat dinner from arising (he has no appetite suppressants available, etc.). It is harder to think of cases in which an agent is unable beforehand to prevent herself from *acting* on a future desire. While Bob may not be able to stave off a desire to eat dinner,

4. The following formulations are somewhat rough. See Mele 1990 for a rigorous analysis of the idea of an irresistible desire (note, however, that Mele does not make the distinctions between such desires that I do in this paper).

Bob can do something now, indeed many different things, to make sure he does not eat dinner and thus cannot act on this desire. For example, he may drive to a remote location where no food is available and is distant enough from anywhere else that he cannot find food until dinnertime has passed.

If we imagine that Bob does not have such *any* methods to prevent *beforehand* his desire from arising or to avoid *beforehand* acting on his future desire to eat dinner (his car is broken down, etc.), then the said desire may be distally irresistible. This is not to say, however, that a distally irresistible desire is also proximally irresistible. It may be that, when the desire to eat dinner arises, Bob is able *at that time* (i.e. at dinnertime) to resist acting on it, even though he could not *beforehand* prevent himself from acting on it.

Permanently irresistible desires are simply desires that are *both* proximally irresistible *and* distally irresistible. A person with OCD who compulsively washes her hands may not be able to prevent her desire to wash her hands arising nor to stop herself acting on it beforehand *or* at the time the desire moves her to action.

Given the above distinctions, there are two natural ways of revising the Simply Irresistible Argument to reflect such disambiguation. This is the first, based on proximally irresistible desires:

1a. Addicts have proximally irresistible desires to take drugs.
2a. If addicts have proximally irresistible desires to take drugs, then addicts are not responsible for taking drugs (when they do so as a result of such proximally irresistible desires).
3a. So addicts are not responsible for taking drugs.

And this is the second, based on permanently irresistible desires:

1b. Addicts have permanently irresistible desires to take drugs.
2b. If addicts have permanently irresistible desires to take drugs, then addicts are not responsible for taking drugs (when they do so as a result of such permanently irresistible desires).
3b. So addicts are not responsible for taking drugs.

A third interpretation involving distally irresistible desires is not natural. If a desire for drugs is distally irresistible but not *proximally* irresistible, then the agent could simply resist taking the drugs at the time (and is thus plausibly morally responsible for taking the drugs). If a desire is distally irresistible *and* proximally irresistible, it is permanently irresistible, which simply takes us back to the second interpretation of the argument.

In what follows I shall argue that once we have disambiguated the original Simply Irresistible Argument, we can show that the revised versions fail. In sections 4 and 5 I explore both interpretations of premise 1 in turn. In section 6, I look at premise 2a (and deal with 2b briefly).

4. Questioning Premise 1a: Addiction and Proximally Irresistible Desires

Premises 1a and 1b attribute proximally and permanently irresistible desires to addicts respectively. 1a, states:

1a. Addicts have proximally irresistible desires to take drugs.

Is this true at least of some addicts? It is relatively natural to think of (severe) addiction as involving such desires, and this view is reflected in various authors' statements on the topic. Consider, for instance, the following:

> …decisions that relate to heroin use are susceptible to powerful physiological and psychological compulsions that usually nullify any semblance of voluntary choice. This is one reason why heroin addicts cannot be considered accountable for their decision to use heroin. (Charland 41, 2002)

> [Addicts] succumb inevitably to their periodic desires for the drug to which they are addicted … these desires are too powerful for him to withstand, and invariably, in the end, they conquer him. He is an unwilling addict, helplessly violated by his own desires. (Frankfurt 1971, 12)

In the above statements, addicts are portrayed as lacking the ability to do other than take drugs because their desires are irresistible.[5] Others, however, are quick to reject premise 1a. For example, Hannah Pickard takes addiction (and other psychopathologies) to involve no such irresistible desires:

> Psychopathology [including addiction]…does not offer us a real case of action without choice between alternatives … there is no compulsion or impossibility of choosing or doing otherwise based on irresistibility of desire. Rather, there is impaired control relative to the norm due to

5. Notably, in Alcoholics Anonymous, addicts are asked to admit that they are powerless over drugs.

> a range of interacting psychological factors and hard choices in difficult life circumstances. (Pickard 2015, 156)

Holton and Berridge think of addiction as involving pathologically intense desires:

> …dopamine works primarily to lay down dispositional intrinsic desires. Addictive substances artifically boost the dopamine signal, and thereby lay down intrinsic desires for the substances that persist through withdrawal, and in the face of beliefs that they are worthless. The result is cravings that are largely outside the control of the addict. (Holton and Berridge 2013, 239)

Even so, they do not hold that such desires are proximally irresistible:

> But this does not mean that addicts are bound to act on such cravings, since they typically retain their faculty of self-control. *The issue is one of difficulty not impossibility*. Controlling an addictive craving is exceedingly demanding. (Holton and Berridge 2013, 239) [italics added]

Some evidence against the thesis that addictive desires are proximally irresistible comes from studies on addicts' sensitivity to a variety of monetary, legal and social incentives (e.g. Higgins et al. 2007; Heil et al. 2008). When faced with such incentives (or threats) many addicts can, at least temporarily, refrain from drug-taking. Further, the very fact that many addicts do recover from their affliction, often spontaneously and without clinical intervention, also provides evidence against addictive desires being irresistible (e.g. Heyman 2009; Foddy and Savulescu 2006). As Pickard further notes:

> If addictive desires are irresistible, and drug-taking and drug-seeking behavior is a direct consequence of a neurobiological disease, then spontaneous recovery and motivated abstinence should be surprising and rare. Yet both are not only possible but common. The natural explanation is that such addicts choose to abstain when they are sufficiently motivated to do so: they are not compelled to use (Pickard 2015, 145)

While I am somewhat sympathetic with the claims of those that deny that addicts have proximally irresistible desires, the evidence they bring to bear does not *conclusively* rule out that a significant number of drug addicts are in fact subject such to proximally irresistible desires. In fact, some indirect evidence supports this thesis, such as the fact

that some addicts knowingly cause themselves great (immediate) physical harm in order to satisfy their desires to take drugs.[6] Further, one may question whether addicts do in fact have the ability to resist addictive desires if they (can) do so only in a restricted number of circumstances, such as when incentivized, etc. (see, for example, Sinnott-Armstrong 2013).

I shall assume, then, that a significant number of addicts *do* have proximally irresistible desires to take drugs. I make this assumption for two main reasons. First, the empirical evidence by no means rules out this hypothesis, and thus it is worth exploring the consequences of addiction on the assumption that 1a is true. Second, the argument I am discussing can be defeated (on both interpretations) even assuming addicts have proximally irresistible desires.

5. Questioning Premise 1b: Addiction and Permanently Irresistible Desires

While the empirical evidence may warrant remaining agnostic about 1a, I shall now present the case for *rejecting* 1b. Premise 1b states:

1b. Addicts have permanently irresistible desires to take drugs.

If addicts *do not* have proximally irresistible desires to take drugs, then neither do they have permanently irresistible desires to do so (since permanently irresistible desires have to be proximally irresistible). Denying 1a, then, commits us to denying 1b. Still, as I mention above, I shall assume that 1a is true—addicts do have proximally irresistible desires. Even on this assumption, 1b is false. Here is my argument for this, based on considerations pertaining to *distally irresistible* desires:

4. Even if addicts have proximally irresistible desires, no (or very few) addicts have distally irresistible desires.
5. If addicts do not have distally irresistible desires, they do not have permanently irresistible desires.
6. Therefore, no addicts (or very few) have permanently irresistible desires (i.e. 1b is false).

Premise 5 follows from the definitions of distally and permanently irresistible desires. This is because permanently irresistible desires are those desires which are proximally *and*

6. For instance, a colleague informed me (in a personal conversation) of an encounter with an addicted individual who burnt his lips so that he could use some crack cocaine.

distally irresistible. Premise 4 is doing most of the work. Why, then, should we accept 4? Why no (or very few) addicts have distally irresistible desires?

First, the evidence supporting the view that addicts have irresistible desires at best indicates they have merely *proximally* irresistible desires. Consider, again, Holton and Berridge's influential account of addiction:

> … the dopamine signals are not learning signals, in the sense that they do not give rise to beliefs, predictions, or memories (real or apparent) at all. Instead, they give rise to desires directly—or, more accurately, to a sensitivity to experience desires when cued with appropriate stimuli. The desire felt is not an instrumental desire, driven by an intrinsic desire for pleasure; instead, it is an intrinsic desire for the drug … (Holton and Berridge 2013, 247)

As we have seen, Holton and Berridge do not think that the desires that addicts gain are (proximally) irresistible. They are rather just *very difficult* to resist. Still, we might easily enough imagine a variation on their view according to which addicts do gain proximally irresistible desires. The pertinent point, as Holton and Berridge emphasize, is that *cues* bring about addicts' pathological desires (such cues may include the presence of the drug, or an addict's drug-dealer, or the environment in which they usually take drugs, etc.). In the absence of such cues, then, addicts are not subject to these desires or, at least, such desires are *less likely* to arise (whether they be proximally irresistible or merely pathologically intense). Because there are many times at which addicts are *not* subject to these desires, these desires are plausibly (distally) resistible at those times. Let us explore this line of argument further.

The strong link between cues and addictive desires is not specific to Holton's and Berridge's theory. For instance, the authors themselves emphasize that the current versions of habit theory suggest that "drugs induce brain systems of action … to form the tendency in the presence of drug cues to perform particular behaviors, behaviors that have been established during previous drug-taking episodes—much like a shoe-tying habit but even more strongly automatic" (Holton and Berridge 2013, 244-5). The importance of cues in addiction is further recognized in clinical practice and the treatment administered to (recovering) addicts which, among other things, focuses on identifying and avoiding such cues (or, alternatively, on one's desensitization to cues). As Sinnott-Armstrong points out, when addicts "face constant drug cues, intrusive thoughts about drugs can resemble obsessions, so many addicts eventually relent or relapse, even if they would not have used drugs in the absence of drug cues" (2013, 128). The importance of

cues in addiction is further evidenced by the fact that people who have previously used drugs are more likely to take up drugs again than those who never started.[7]

Addicts are not exposed to such cues all the time. Given that cues do play a crucial role in triggering addictive desires, addicts will not have these desires at all times (even though they may still have *dispositions* to gain such desires). There being (possibly extended) blocks of time when addicts are not subject to relevant cues leaves them enough room to implement numerous methods that can stave off any proximally irresistible desires which might arise in the future. For instance, addicts may take steps to avoid their dealers, acquaintances who also take drugs, and situations and places in which they usually take drugs. Alternatively, they may take steps to ensure that, even if they gain addictive desires, they will not be able to act on them; for example, they can give their money to someone else so that they cannot purchase more drugs. Such methods are not esoteric—upon reflection people can easily come up with such ideas—and thus they are epistemically accessible to addicts. To truly find a case in which an addict's desire really is distally irresistible, the agent must be unable to apply any effective and epistemically accessible method to prevent her gaining or acting on a potential future desire. Such cases will be, if not nonexistent, at least exceedingly rare.

That addicted individuals have a number of such methods available to them is further supported by the fact that obtaining drugs and fulfilling addictive desires is a rather complex process that requires a good deal of organization and planning. It is a well-known fact that addicts often go to extraordinary lengths in order to satisfy their addictive desires, be it with regards to obtaining means to secure drugs, securing the drugs themselves, finding a suitable location or time for using and otherwise creating opportunities for drug-use. Given that the drug-seeking behavior is often temporally-separated from the drug-taking behavior, and involves multiple steps, this gives addicts plenty of opportunities to intervene at many junctions along the way.

To illustrate the point, imagine a case in which Bob, a cocaine user, knows (given his previous experiences) that he will want to use drugs sometime this week. Bob will have to do, minimally, two things: obtain the drugs and create an opportunity for using the drug (often addicts will have to plan for both of these things, and, even more frequently, for at least one of them). Both actions involve a series of what may be rather complex steps: calling the supplier, meeting up with the supplier, securing enough money for the transaction, and securing a suitable environment, etc. Each of these steps can be

7. For some interesting discussion on cues see, for example, Robins and Slobodyan 2003; Hyman and Malenka 2001, Carter and Tiffany 1999.

broken down to even more sub-steps. The relevant drug-seeking behaviors are thus quite complex, requiring careful guiding and sustaining. This provides Bob with opportunities to refrain from these behaviors at numerous points (this contrasts with cases in which Bob's access to drugs is easy and immediate, in which case he will find it much harder, if not impossible, to resist).

Addicts, then, often have the ability and opportunity to prevent themselves from acting on addictive desires (either by stopping these desires from arising, or by blocking their effectiveness). Still, one might object that having such physical ability and opportunity is not enough—addicts must also be able to be *sufficiently motivated* to take appropriate countermeasures. While motivation is not sufficient for putting such measures in place, it is arguably necessary. At first blush, this seems unproblematic: the fact that the occurrence of addictive desires seems to be largely tied with the relevant cues suggests that there will be numerous occasions on which addicted individuals are able to be sufficiently motivated.

However, in order to be motivated to refrain from taking drugs, one must *think about* taking drugs (as something to avoid). But merely *thinking* about drugs can give rise to addictive desires. This may, given the significant motivational strength of such desires, diminish one's motivation for putting relevant countermeasures in place. Perhaps, then, a significant number of addicts can never be sufficiently motivated not to take drugs, and thus their desires for the drugs remain distally irresistible.

A few points need to be mentioned in relation to this. It is, indeed, rather plausible that merely thinking about drugs can give addicts desires for the drug. However, not all cues will have equally strong effects on addicts. Seeing the drug or being offered the drug, for example, will likely impact an addict a lot more than merely reading about it in the papers. So, even if addicts sometimes gain irresistible desires, they may not do so in response to *all* the relevant cues (some cues might give addicts only resistible but still extremely strong desires). Further, even if some addicts may gain proximally irresistible desires to use just by thinking about drugs, it is unlikely that these desires will also be distally irresistible. If addicts gain proximally irresistible desires to use drugs but cannot immediately satisfy such desires, their motivation to use drugs will likely decrease, as is often the case with desires not involved in addiction.[8] This would then allow addicts to be sufficiently motivated not to take drugs.

8. The fact that a desire may be irresistible at a time does not mean that it will always be irresistible or that one will have such a desire manifesting up until it is satisfied. Persistence of desires as well as persistence of their motivational strength is sensitive to many factors, such as whether it is possible, or how easy it is,

6. Questioning Premise 2a and Premise 2b: Irresistible Desires and Moral Reponsibility

I have argued that addicts do not have distally irresistible desires. If I am right, then the second interpretation of the Simply Irresistible Argument fails. If any interpretation is to succeed, then, it must be the first.

I have already granted the truth of 1a. That is, I am happy to assume that addicts have proximally irresistible desires. In what follows I shall argue that, given this very assumption, 2a is false (or, at the very least, implausible and unsupported). To recap, premise 2a says:

2a. If addicts have proximally irresistible desires to take drugs, then addicts are not responsible for taking drugs (when they do so as a result of such proximally irresistible desires).

My argument against 2a runs as follows:

7. Addicts have proximally irresistible desires to take drugs (assuming 1a to be true).
8. Any proximally irresistible desires (to take drugs) are not distally irresistible.
9. If addicts' proximally irresistible desires are not distally irresistible, then addicts *are* responsible for taking drugs (even when they do so as a result of such proximally irresistible desires to take drugs).
10. Therefore, addicts have proximally irresistible desires AND addicts are responsible for taking drugs (even when they do so as a result of these proximally irresistible desires).

Premise 7 I am granting for the sake of argument. If it is false, then 1a is false and the argument I am criticizing falls at the first hurdle. Premise 8 is entailed by premise 4, which I have defended above. In essence, addicts not subject to cues do not have pathologically strong or irresistible desires, and have many opportunities and methods to prevent themselves gaining such desires or acting on any such desires that may arise. The conclusion, 10, entails that 2a is false. Premise 9 is what I have left to support.

The basic idea behind 9 is that, though after a certain time the addict may not able to do other than take drugs as a result of her irresistible desire to do so, she *did* have, at some point in the past, the ability and opportunity to prevent herself gaining this desire

to satisfy these desires. However, for a desire to be truly permanently irresistible, it would have to remain irresistible at all times at which one has such a desire (and at all times *before* one has it).

or to prevent herself acting on it. Thus she could have done other than end up taking the drugs. And if she could have done other than take drugs, she had sufficient control over whether she did so or not. It's simply that she had this control earlier than when she did take the drugs. Such control still suffices for moral responsibility (bracketing any other worries about moral responsibility that are unrelated to irresistible desires).

To bring this idea out, consider the following fantastical example:

> Oz is a werewolf. On a full moon, he transforms and, if not properly bound, rampages through the streets of Sunnydale and kills people. When transformed, Oz acts on a proximally irresistible desire to kill. Oz knows all this, knows a full moon is coming, and has the ability now to chain himself up to stop him acting on such a future desire.

If Oz fails to chain himself up, he is clearly morally responsible for any killing he does. Though he may not have control over his actions at the time he kills someone, the control he has beforehand more than suffices for his being morally responsible. Any remorse, regret and guilt Oz might feel is perfectly appropriate as are feelings of indignation and resentment from others towards him.

The same lesson straightforwardly applies to addicts. If they have control at certain points beforehand over their drug-taking on any individual occasion, they are morally responsible for taking drugs (on that occasion).[9] We cannot yet infer, however, that they are *blameworthy* for taking drugs—on many occasions it may be the appropriate thing to do (perhaps because, on such occasions, the methods of preventing themselves taking the drugs are too costly, or unethical). Still, it is likely that such behavior often *is* blameworthy, i.e. when the costs of preventing themselves acting on the drugs are not so costly, and not unethical.

One could perhaps object that *local* control, i.e. control that an agent has in any given moment over her immediate actions, is somehow more relevant to responsibility than *distal* control, i.e. control over one's future actions. However, I see no reason why that ought to be the case, and without such a reason, one cannot dismiss addicts' distal control as being less relevant than local control (or completely irrelevant). It is often the case that local control *requires* and is *enabled by* distal control. Take implementation intentions for instance (e.g. Gollwitzer 1999). Implementation intentions are plans of the form "if (or when) X obtains, then I will Y", and the execution of these intentions is based on a cue which the agent specifies beforehand. Once the cue is encountered, the agent

9. This does not require that they have control over not getting addicted in the first place.

responds as specified in her intention. In such cases, the agent's local control of her action requires previously gaining the implementation intention. By forming such an intention the agent, distally controls her action.

The above considerations do not apply to Premise 2b which says that agents should be exculpated for their behavior if it follows from *permanently irresistible* desires. In fact, this premise seems quite plausible. As we have already briefly discussed, if an agent cannot, at any time at all, do anything to refrain from acting in a certain way (and she does what she does only *because* she cannot do otherwise), then it seems problematic to hold her responsible for her action given that she could do no other. (Again, this presupposes that responsibility requires some sort of control; a non-volitionist might simply bite the bullet here.) However, even if we accept 2b, the second version of the argument will not go through since we have already rejected premise 2a.

In conclusion, the Simply Irresistible Argument fails on both natural interpretations. Addicts are morally responsible for taking drugs even if they act on proximally irresistible desires. The Simple Irresistible Argument is resistible.

7. Moral Responsibility, Blameworthiness and Patterns of Behavior

When assessing whether, or to what degree, addicts are morally responsible (or blameworthy) for their drug-taking behavior, it is not sufficient to consider the amount of control addicted individuals have over this kind of behavior on individual occasions. One also ought to consider how much control addicts have over their drug-related behavior over *an extended period of time*. Even if one has control over (a number of) individual actions, this does not mean that one also has control over a *pattern of behavior* made up of such individual actions (including omissions to act in certain ways). Control over a pattern of behavior amounts to being able to intentionally engage in or refrain from the relevant behavior on a *sufficient* number of instances. On this understanding, then, having control over a drug-taking pattern of behavior requires that one has control over a sufficient number of instantiations of this pattern; that is, one has control over (not) taking drugs on different occasions for an extended period of time.

Now even if addicts do not have (proximally or permanently) irresistible desires to use drugs on any individual occasion, and their control over individual actions is thus not eradicated as a result of irresistible desires, their control over drug-related behavior in general may still be severely diminished, or even non-existent. While the views on the exact nature of addictive desires and whether they are irresistible diverge, most would agree that, minimally, addictive desires have *great* motivational strength and are very

difficult to resist. This may present a problem for successfully resisting addictive (or any other strong) desires on a regular basis. Studies on the phenomenon of *ego-depletion* show that resisting wayward desires, among other things, temporarily impairs one's self-regulation capacities (e.g., Baumeister et al. 1998). The ego depletion data show that self-controlled (and some other) behaviors draw on a limited resource which can be used up and which takes time to be replenished again: "the self's acts of volition draw on some limited resource … and that, therefore, one act of volition will have a detrimental impact on subsequent volition" (Baumeister et al. 1998, 1252).

Continually resisting strong desires is likely to be rather depleting, leaving an agent with fewer or limited resources to fend off future temptations. So even if addictive desires are not irresistible, repeatedly resisting these desires will likely result in suboptimal availability of the resource(s) used in self-regulation (*cf.* Levy 2006). Having control over any or any number of individual instantiations of a general pattern of behavior thus does not straightforwardly amount to having control over such pattern. Control over the latter requires adjusting to and accommodating for possibly rather significant depleting effects of previous self-regulation. Some patterns of behavior, keeping everything else equal, are certainly easier to control than others. Behaviors which do not require significant amounts of effort will be easier to upkeep than those which do; a pattern of briefly scanning one's email every morning will be easier to control than that of resisting to have a drink. Regardless, then, of how successful one may be in controlling one's addictive desires on individual occasions, it is far from obvious that one has sufficient control over (not) acting on such desires on a long-term basis.

What of the implications for moral responsibility? One likely consequence is that being responsible for an individual action (or omission) does not amount to being responsible for the corresponding pattern of behavior. If we ground responsibility for individual actions in the degree of control that one has over these actions, then, plausibly, considerations about responsibility with regards to long-term patterns of behavior need to be sensitive to considerations about control over these behavioral patterns as well, in a parallel way. Then, given that one's control over a long-term pattern of behavior might be diminished due to the factors discussed above, this should also be reflected in our moral appraisal of the agent. An agent's responsibility over her general drug-taking behavior thus may be diminished—*even if* we hold her responsible her behavior on individual occasions.

8. Conclusion

My debate above concerns substance addictions. It is worth pointing that the arguments I consider can be rehashed to make parallel claims about *behavioral* addictions; to the effect that subjects with such addictions have corresponding irresistible desires, and are thus not responsible for acting on these desires. However, such arguments would be arguably even less plausible than those concerning substance addictions. This is because drug addictions are typically thought to involve the *strongest* addictive desires. With less strong desires, such as those involved in behavioral addictions, it is even less convincing that these desires are irresistible, rendering agents unable to refrain from the relevant behavior. If individuals with behavioral addictions are then to be absolved from responsibility for fulfilling (or attempting to fulfil) their addictive desires, this cannot be grounded in considerations about the strength of such desires.

One could also alter the above arguments to include behaviors wider than drug-taking; namely various types of *drug-seeking* behavior. Again, such arguments would be arguably less plausible than those concerning drug-taking. Desires to use drugs, whatever these desires are exactly for or about (be it the drug itself, pleasure, etc.), are stronger than drug-related desires which one might gain to help fulfill desires to take drugs (desires to obtain drugs, etc.). First, such instrumental desires most likely arise differently than desires for drugs (which are likely a result of pathological processes and aberrations). Second, drug-related instrumental desires are highly controllable in various ways; for instance an agent with such desires is responsive to various practical considerations. Drug-related instrumental desires seem to then have significantly less motivational strength than desires to take drugs. It is very unlikely that drug-seeking desires should be irresistible (permanently or even proximally), and therefore, they cannot provide a basis for exculpating addicts for their drug-seeking behavior.

While I believe that the arguments I have presented, concerning substance addictions and drug-taking behavior, are perhaps the most viable forms of arguments involving addicts' irresistible desires (to the conclusion that addicts are not responsible for acting on their addictive desires), I also think that both arguments ultimately fail. This is not to say that addicts *are* in fact responsible (or blameworthy) for taking drugs. However, if they are not so responsible (or blameworthy), this cannot be due to the motivational strength of their addictive desires.

References

Baumeister, Roy. F, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology* 74 (5): 1252–65.

Carter, B. L., and S. T. Tiffany. 1999. "Meta-analysis of cue-reactivity in addiction research." *Addiction* 94 (3): 327–340.

Charland, Louis C. 2002. "Cynthia's dilemma: consenting to heroin prescription." *The American Journal of Bioethics* 2 (2): 37–47.

Frankfurt, Harry G. 1969. "Alternative Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 829–839.

Frankfurt, Harry G. 1971. "Freedom of the will and the concept of a person." *The Journal of Philosophy* 68 (1): 5–20.

Fischer, John Martin. 2010. "The Frankfurt Cases: The Moral of the Stories." *Philosophical Review* 119 (3): 315–336.

Foddy, Bennett, and Julian Savulescu. 2006. "Addiction and autonomy: can addicted people consent to the prescription of their drug of addiction?" *Bioethics* 20 (1): 1–15.

Gollwitzer, Peter M. 1999. "Implementation Intentions: Strong Effects of Simple Plans." *American Psychologist* 54 (7): 493–503.

Heil, Sarah H., Stephen T. Higgins, Ira M. Bernstein, Laura J. Solomon, Randall E. Rogers, Colleen S. Thomas, Gary J. Badger, and Mary Ellen Lynch. 2008. "Effects of voucher-based incentives on abstinence from cigarette smoking and fetal growth among pregnant women." *Addiction* 103 (6): 1009–1018.

Heyman, Gene. 2009. *Addiction: a Disorder of Choice*. Cambridge MA: Harvard University Press.

Higgins, Stephen T., Sarah H. Heil, Robert Dantona, Robert Donham, Martha Matthews, and Gary J. Badger. 2007. "Effects of varying the monetary value of voucher-based incentives on abstinence achieved during and following treatment among cocaine-dependent outpatients." *Addiction* 102 (2): 271–281.

Hyman, Steven E., and Robert C. Malenka. 2001. "Addiction and the Brain: The Neurobiology of Compulsion and its Persistence." *Nature Reviews Neuroscience* 2 (10): 695–703.

Holton, Richard, and Kent C. Berridge. 2013. "Addiction Between Compulsion and Choice." In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, edited by Neil Levy, 239–265. Oxford: Oxford University Press.

Levy, Neil. 2006. "Addiction, Autonomy and Ego-Depletion: A response to Bennett Foddy and Julian Savulescu." *Bioethics* 20 (1): 16–20.

Mele, Alfred R. 1990. "Irresistible Desires." *Noûs* 24 (3): 455–472.

Mele, Alfred R. 2014. "Self-control, Motivational Strength, and Exposure Therapy." *Philosophical Studies* 170 (2): 359–375.

Pickard, Hannah. 2015. "Psychopathology and the Ability to Do Otherwise." *Philosophy and Phenomenological Research* 90 (1): 135–163.

Robins, Lee Nelken, and Sergey Slobodyan 2003. "Post-Vietnam heroin use and injection by returning US veterans: Clues to preventing injection today." *Addiction* 98 (8): 1053–60.

Sinnott-Armstrong, Walter. 2013. "Are Addicts Responsible?" In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, edited by Neil Levy, 122–142. Oxford: Oxford University Press.

Smith, Angela M. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 (3): 367–392.

Journal of Cognition and Neuroethics

Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism

Derek Jones
University of Evansville

Biography

Derek Jones is Assistant Professor of Philosophy and Director of Cognitive Science at the University of Evansville. He works on topics at the intersection of the philosophy of mind, action and biology and is currently writing a book on the biological foundations of agency.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Jones, Derek. 2015. "Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism." *Journal of Cognition and Neuroethics* 3 (1): 217–228.

Agency through Autonomy: Self-Producing Systems and the Prospect of Bio-Compatibilism

Derek Jones

Abstract
In this paper I motivate a biologically-oriented compatibilism that is consistent with Daniel Dennett's compatibilist account but which avoids some of the recent criticism directed towards it, specifically challenges to his "mild realism" and his reformulation of the principle of alternate possibilities. I argue that a theory of free will that grounds agency in the dynamics of autonomous self-producing systems can show the ways in which agents may engage with and contribute to a given past in uniquely agential ways.

Keywords
Adaptivity, agency, autonomy, autopoiesis, compatibilism, complex systems, Dennett, Jonas, Kant, Lorenz, process biology

Introduction

Determinism is the thesis that past events together with the laws of nature fully determine future events; in a deterministic universe there is exactly one possible future at any given time. One formulation of the problem of free will is that if determinism is true, then our actions are not truly "up to us"—any causally-efficacious state within us would itself have been fully determined by some prior cause, which itself would have been determined, and so on. Given the transitivity of the determination relation, it follows that the initial configuration of the universe, together with its laws, fully determines the final configuration of the physical universe. Anything obtaining between those events is simply along for the ride.

However, there is reason to think that varieties of freedom may nonetheless arise between the birth and heat death of a deterministic universe. Hans Jonas (1966) suggests that such varieties are distinctly biological:

> ... it is in the dark stirrings of primeval organic substance that a principle of freedom shines forth for the first time within the vast necessity of the physical universe—a principle foreign to suns, planets and atoms.

In what follows I sketch how this biological "principle of freedom" may emerge in a deterministic universe. My focus is not on capacities required for morally significant free will, but on the necessary conditions for being a free agent of any kind. It will help to begin with a famous case of naturalistic compatibilism.

Dennett's Compatibilism

In *Freedom Evolves* Daniel Dennett suggests a way in which free will worth wanting might arise in a deterministic universe. He describes creatures in a deterministic "Game of Life"-style toy universe that, upon achieving an appropriate degree of complexity, are best described in behavioral language—they "seek," "avoid," "eat" and so on. These terms describe the systems' *capacities*. To say that a creature in this world *can* avoid harm is to say that it is organized in such a way that it *will* avoid harm in a certain range of conditions. It exercises its capacities for avoidance when it *does* avoid harm. If we were to restart its universe a million times it might avoid harm in precisely the same way each time, but this fact in no way robs the organism of its abilities to avail itself of the opportunities presented by its world (Dennett 81).

Some might not wish to describe Dennett's creature as *freely* avoiding. When we claim that an agent can act freely, we mean that it was possible that they could have done something other than what they did. Genuine freedom involves the agent's ability to collapse a range of possible futures into a single actual event. This notion, known as the principle of alternate possibilities (PAP), has been challenged by Frankfurt (1969) and others, but Dennett accepts it, choosing instead to blunt the challenge by distinguishing between wide and narrow conceptions of possibility.

The narrow reading of "Pat could have Φed" suggests that, given the fixed past up to the time T of the action, Pat could have either Φed or not Φed at T. The wide reading of "Pat could have Φed" suggests that, had the universe been different in some way prior to T, Pat would have Φed at T. Dennett argues that the wide reading underlies most of our empirical tests of causal power. For example, we confirm that one *could* have sunk a missed putt in golf by repeating the putt in circumstances *similar* to those obtaining during the original putt. Performance in *identical* circumstances is irrelevant to the investigation.

This wide reading is entirely consistent with determinism. If what we are saying when we claim that Pat could have Φed is simply that, had previous conditions been different, Pat would have Φed, we are identifying a range of possible deterministic unfoldings of the universe (individuated either by starting conditions or laws) that

happen to include Pat. The agent's causal powers cash out in terms of how competently it copes with whatever unfolding it happens to face. If Pat could only successfully Φ in one or two of the various relevant possible timelines leading up to T, we might judge them as being less competent than an agent that Φs across a broader range. In some cases we might be warranted in chalking the performance up to luck. Competence—what the agent *could* do—amounts to facts about the agent's organization and how robustly it copes with its environment.

John Martin Fischer (2003) agrees that there is a place for the wide reading of possibility but rejects Dennett's claim that it is the only reading that matters to "serious investigators of possibility." When we say we could have done otherwise we do not typically think that we are referring to alternate starting conditions of the universe. Rather, we believe that freedom "consists in [one's] power to add to the given past, holding the natural laws fixed" (635). The relevant sort of additive power goes beyond mere contribution. If lightning strikes a tree, igniting it and causing a forest fire, that tree does contribute to the given past—there would have been no forest fire had it not existed. Moreover, the tree's contribution depends on one of its dispositional properties (flammability), which it possesses in virtue of its physical organization. Still, the tree is not a free agent. Fischer worries that compatibilism cannot advance if it fails to at least *respect* libertarian intuitions about what freedom entails. The compatibilist may reject the standard interpretation of PAP, but in doing so they assume the burden of showing how free agents contribute to the given past in distinctly agential ways.

A related worry targets Dennett's mild realism. Dennett argues that we find the language of agency indispensible once a system achieves a certain level of complexity, but that it is a mistake to look for anything metaphysically deeper than that. But our intuitions about freedom include the idea that agents have some privileged status in the causal order of things, independent of our interpretive practices. Even if the agent is not the *ultimate source* of its behavior, it has objective properties that allow it to contribute to the progression of events in uniquely agential ways. In what follows I offer a view of agency that is compatible with Dennett's view but that privileges the role of the agent in a way that does (more) justice to our intuitions about free will.

Primitive Agency

Most discussion of agency emphasizes higher-order processes of deliberation and planning. Such discussions are valuable, but there is reason to think that a bottom-up approach to understanding action may be equally illuminating. Frankfurt (1978) argued

that a complete action theory would accommodate an active/passive distinction in animals that are incapable of deliberate action. To use his example, there is a difference between when a spider moves its leg and when its leg is moved from without. Tyler Burge (2009) argues that very basic systems such as eukaryotic cells count as primitive agents. Despite lacking capacities required for deliberative agency, there is an intuitively plausible distinction both between things those organisms do and things that happen to them, and between things organisms do and things their parts do—to use Burge's example, the *amoeba* eats but *its gullet* digests. Burge characterizes primitive agency as whole-organism functional behavior, but it is far from clear when to characterize a behavior as "whole-organism," particularly in very simple systems.

Biological interest in the whole organism as an object of study has recently surged (for a helpful summary see Nicholson 2014), but it has a long History. Kant (1987/1790) acknowledged the uniqueness of organic life in his *Critique of Judgment,* noting that their parts "are reciprocally cause and effect of their form" and that "the possibility of [the system's] parts… [must] depend on their relation to the whole" (287). Konrad Lorenz (1996/1944) defined organisms as organic *entities*, which he defined as "regulatory systems of universal, reciprocal causal connections" (137). Entities are not mere constructions of their parts because the activities of those interdependent parts are subordinated to the activities of whole entities—the parts of living systems are continually changing, and their changes are governed by the constitution and activities of the systems they comprise. Moreover, since life depends on a continuous process of endothermic assimilation and exothermic dissimilation of matter—the living system persists by breaking down its parts and rebuilding them—the whole entity displays greater invariance than its parts (85).

Recent work on self-organization offers a framework for understanding whole-system behavior. Alicia Juarerro (1999) suggests that agent-individuation amounts to identifying the proper collective variables governing the behavior of a complex system—we are "eddies of order" (145). Whole-organism behavior might be best described as those processes that correspond to changes in the order parameter values. However, it is far from obvious what ought to guide the process of order parameter selection. Indeed, from the standpoint of the complexity theorist the matter of what systems count as agents and which do not may be difficult to settle objectively—in a world of flux, boundaries may be drawn as the observer sees fit (Dennett's "mild realism" may be motivated by such considerations). Moreover, not all self-organizing systems are agents; storm clouds, tornadoes and crystal lattices cannot act. More work must be done to find agency worth having.

From Order to Value

One plausible move is to distinguish the living organism as a locus of purpose or *value*. Kant refers to self-organizing systems as "natural purposes." Lorenz distinguishes organisms from "physical gestalts" by their *finality* "in the sense of purposive survival value" (142). Jonas (1966) argues that purposiveness and value arise from the "needful freedom" of the organism, engendered by the very metabolic processes that differentiate it from its environment.[1] Jonas characterizes the phylogeny of organismic life as a series of systems that enjoy increasing freedom from their environment as they increase in complexity. The earliest form of freedom manifests in the self-organizing system's apparent violations of the Second Law of Thermodynamics—the free system first and foremost "oppos[es] in its internal autonomy the entropy rule of general causality" (5).

Jonas argues that living systems are unique among self-organizing systems in that they are essentially concerned with self-production,[2] a process through which the system distinguishes itself as *autonomous*. Definitions of biological autonomy vary widely,[3] but most characterize it as a property of far-from-equilibrium, operationally closed, dissipative self-organizing systems. This characterization can be made more concrete by examining a paradigmatic case of biological autonomy: the autopoietic system.

Maturana and Varela (1973) define living systems as autopoietic machines:[4]

1. Here Kant and Jonas break with Lorenz, who offers an evolutionary teleofunctional approach. This disagreement has no bearing on the present line of argument and will not be addressed here.

2. … living things… are unities of a manifold, not in virtue of a synthesizing perception whose object they happen to be, nor by the mere concurrence of the forces that bind their parts together, but in virtue of themselves, for the sake of themselves, and continually sustained by themselves… *This active self-integration of life alone gives substance to the term "individual."* (Jonas 1966, 79 [my emphasis]).

3. Ruiz-Mirazo and Moreno (2004) define *basic autonomy* as "the capacity of a system to manage the flow of matter and energy through it so that it can, at the same time, regulate, modify, and control: (i) internal self-constructive processes and (ii) processes of exchange with the environment." (240). Thompson (2007), citing Varela (1979), defines the autonomous system as a system whose constituent processes "(i) recursively depend on each other for the generation and their realization as a network, (ii) constitute the system as a unity in whatever domain they exist, and (iii) determine a domain of possible interactions with the environment." (44). Hooker (2011) defines autonomy as "the internally organized capacity to acquire ordered free energy from the environment and direct it to replenish dissipated cellular structures, repair or avoid damage, and to actively regulate the directing organization so as to sustain the very processes that accomplish these tasks" (35).

4. Here "machine" simply denotes systems that are defined by their organizations.

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network. (79)

This process amounts to the self-production of the system. Crucially, the autopoietic system forms and continuously maintains a boundary, distinguishing its internal processes of the system from those of its environment. The boundary is both the product of and a necessary condition for the cell's metabolism, simultaneously limiting, contributing to and being sustained by the system's internal dynamics. Through this process of self-production and differentiation the system distinguishes itself from its environment as an autonomous unity.

The living system's "needful freedom" is due to the fact that its apparent violation of the Second Law is only apparent: it cannot remain in a far-from-equilibrium state without energy from its environment. Paradoxically, it cannot differentiate itself from its environment without continuously engaging it. Here the system's boundary serves to distinguish organization-sustaining elements from harmful elements. Furthermore, structure of the organism creates what Sørensen and Zienke (forthcoming) call an "asymmetry of normativity": depending on the structure of the system at a given time, certain features of its environment contribute to its self-maintenance—and thus are good for it—and others do not. In this way the organism's organization defines its subjective world—what the ethologist Jakob von Uexküll called its *Umwelt*.

This interaction between agent and environment evokes Merleau-Ponty's (1963) metaphor of "a keyboard which moves itself in such a way as to offer such or such of its keys to the in itself monotonous action of an external hammer" (13). Value is not an objective feature of the environment (as it might be if the structure of the musical piece were found in the environment for the passive keyboard to receive). It is constructed by the system as it navigates its environment over time. The system and its environment generate meaning through their mutual interactions at the system's boundary. Autonomous agents are not mere "eddies of order," but rather wellsprings of value.

Embedded Norms

The normative structure of even basic agency extends beyond the matter of maintaining one's organization over time. Jonas (1966) argues that the norms governing animal movement are unique within the organic world. Unlike non-motile organic systems, which either exploit the materials with which they are in direct contact or die, motile animals evolved to fit an environment wherein the materials needed for survival are spatiotemporally distant. Unlike certain plants, which can generate the materials they need to live by exploiting light energy and minerals from their soil, animals cannot manufacture the proteins, carbohydrates, fats, etc., they need on their own. So organized, animals both *must* and *can* seek out these materials in other organisms. We might say that their organizations *dynamically presuppose* this need.[5]

Jonas views this shift from basic metabolic need to what he calls *appetite* as a gradual increase in the "transcendence" of life beyond its "point-identity" (85). This "transcendence" involves a sort of dynamic presupposition along both spatial and temporal dimensions: the sensorimotor organism depends upon an energy supply that lies well beyond its boundary. Jonas notes that one consequence of this gap is a corresponding gap separating "action from its purpose" (104). Motile animals must perform "intermediate" movements, which contribute to metabolism only indirectly. These movements draw upon the organism's energy reserves, "an expenditure to be redeemed only by [its] eventual success" (ibid). Thus animal agency essentially involves gambling with one's energy reserves in hopes that the environmental payoff will have been worth the risk, for example, by funneling them into pursuit or avoidance behaviors. Stimuli are not merely "good" or "bad," but also "worth it" or "not worth it" from the perspective of a sensorimotor agent with real-time energetic needs in an uncertain environment.[6] I submit that the enaction of these "embedded" norms is the hallmark of primitive agency.

---

5. "That is, it is derived from the norm of contributing to the maintenance of the conditions for the far from equilibrium continued existence of the system… More generally, a process dynamically presupposes whatever those conditions are, *internal to the system or external to the system*, that support its being functional for the system." Beer (forthcoming) refers to the agent as prospering in "precisely those environments to whose spatiotemporal structure its autopoietic dynamics is matched" (28). It is this "match" or "fit" between agent and environment that I mean to capture with this term, applied in a different context in Bickhard (1993).

6. Here a connection can be drawn with Millikan's (1993) discussion of the embedded character of functional behavior. Millikan argues that the difference between functional behavior and other functional state changes is determined by whether the state change effects changes in the organism's environment such

How does any of this give us "freedom worth having?"

This approach motivates a biologically-oriented compatibilism that is friendly to Dennett's approach but avoids some of its shortcomings. Biologically autonomous agents are composed of matter that obeys the deterministic laws of nature. However, as a complex system the autonomous agent's causal contributions to the world are distinct from those of other objects. This is because complex systems persist by imposing constraints on their constituent parts. Atoms "caught up in the life of an organism" (Van Inwagen 1990) behave in ways they otherwise would not—their individual freedom is restricted by the system's global structure. Entrainment is common to complex systems—the vortex of water that emerges when one drains a bathtub—an *actual* "eddy of order"—is a clear case.

But as Van Inwagen notes, the living system does not "simply deposit and withdraw sequentially an invariant sum of energy" as an actual eddy might but rather "takes the energy it finds and turns it to its own purposes" (89). The world proceeds the way it does in part because the autonomous agent has the needs it does; the processes that constitute the agent's perspective and that engender its needs are doing the relevant constraining. The biologist J.Z. Young offers an apt characterization:

> The essence of a living thing is that it consists of atoms of the ordinary chemical elements... caught up into the living system and made part of it for a while. The living activity takes them up and organizes them in its characteristic way. The life of a man consists essentially in the activity he imposes upon that stuff. (1971, 86–87)

Thus the organization of a living system entrains its constituent matter into activities that serve its perspective and satisfy its needs. I have argued that in the case of the sensorimotor agent, those needs primarily involve the investment of energetic resources in adaptive sensorimotor activity. They may motivate energetic investment in pursuit behaviors at some times and avoidance at others. But it is *up to the organism* how its resources are invested—other organisms with distinct structures might have behaved quite differently in the same circumstances.

We may now revisit Dennett's claims about what an organism could have done. The autonomous agent's organization at any given time determines the range of

---

that the organism gets a return on its investment. This is why, for example, the clam's slowing its activity in cold water does not count as behavior but the spider's pursuit behavior does: only the latter involves an energetic *investment*, rather than a mere expenditure.

environmental features to which it will be receptive (effectively determining the range of possible past timelines that will matter for its action). Its organization also determines the trajectory of future causal events: the agent is successful when it is able to steer that trajectory in ways that satisfy its needs; it fails when it cannot. We can judge the agent's capacities by appeal to its robustness across possible timelines: the range of alternate causal trajectories that it can steer in its favor. To say that the rabbit could have avoided the hawk is to suggest that there is a range of possible deterministic unfoldings of the universe that happen to include that rabbit at that time, and that in some of those unfoldings the rabbit's organization successfully channels enough energy into the task of avoidance.

None of this denies our intuition that freedom consists in a distinctly agential power to add to a given past. The processes that channel matter and energy through the organism operate as they do *because* of the organism's perspective and needs. The agent may not be the *ultimate* source of its behavior—this is, perhaps, too much to ask—but it does matter in a distinctly agential way. I submit that this is a form of free will worth having.

References

Beer, Randall. 2004. "Autopoiesis and Cognition in the Game of Life." *Artificial Life* 10 (3): 309–326.

Bickhard, Mark. 2004. "The Dynamic Emergence of Representation." In *Representation in Mind: New Approaches to Mental Representation*, edited by H.Clapin, P. Staines and P. Slezak, 71–90. Oxford, UK: Elsevier Inc.

Burge, Tyler. 2009. "Primitive Agency and Natural Norms." *Philosophy and Phenomenological Research* 79 (2): 251–278.

Dennett, Daniel. 2003. *Freedom Evolves*. New York: Viking.

Di Paolo, Ezequiel. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4 (4): 429–452.

Fischer, John. 2003. "Review of Freedom Evolves by Daniel C. Dennett." *Journal of Philosophy* 100 (12): 632–637.

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66 (23): 829–39.

Frankfurt, Harry. 1978. "The Problem of Action." *American Philosophical Quarterly* 15 (2): 157–162.

Hooker, Cliff. 2011. *Handbook of the Philosophy of Science. Volume 10: Philosophy of Complex Systems.* Waltham, MA: Elsevier B.V.

Inwagen, Peter. 1990. *Material Beings*. Ithaca, N.Y.: Cornell University Press.

Jonas, Hans. 1966. *The Phenomenon of Life: Toward a Philosophical Biology*. New York, NY: Harper & Row Publishers, Inc.

Juarrero, Alicia. 1999. *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge: MIT Press.

Kant, Immanuel. 1987. *Critique of Judgment.* Trans. W.S. Pluhar. Indianapolis, IN: Hackett Publishing Company.

Lorenz, Konrad. 1996. *The Natural Science of the Human Species: An Introduction to Comparative Behavioral Research. The "Russian Manuscript" (1944–1948)*. Cambridge: MIT Press.

Maturana, Humberto, and Francisco Varela. 1980. *Autopoiesis and Cognition: the Realization of the Living.* Dordrecht: D. Reidel Pub. Co.

Merleau-Ponty, Maurice. 1963. *The Structure of Behavior*. Boston: Beacon Press.

Nicholson, Daniel. 2014. "The Return of the Organism as a Fundamental Explanatory Concept in Biology." *Philosophy Compass* 9 (5): 347–359.

Ruiz-Mirazo, Kepa, and Alvaro Moreno. 2004. "Basic Autonomy as a Fundamental Step in the Synthesis of Life." *Artificial Life* 10 (3): 235–259.

Sørensen, Mikkel, and Tom Zienke. Forthcoming. "Agents Without Agency?" *Cognitive Semiotics*.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Belknap Press.

Varela, Francisco, Evan Thompson, and Eleanor Rosch. 1992. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Young, John. 1971. *An Introduction to the Study of Man*. Oxford: Oxford University Press.

Journal of Cognition and Neuroethics

The Limits of a Pragmatic Justification of Praise and Blame

Ryan Lake
Clemson University

Biography

Ryan Lake earned his PhD in Philosophy from the University of Miami in 2013. His main research interests are in ethics and metaphysics, in particular the problem of free will and moral responsibility, and related issues regarding social policies and criminal justice. He is currently teaching as a full-time Lecturer at Clemson University.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Lake, Ryan. 2015. "The Limits of a Pragmatic Justification of Praise and Blame." *Journal of Cognition and Neuroethics* 3 (1): 229–249.

The Limits of a Pragmatic Justification of Praise and Blame

Ryan Lake

Abstract
In recent decades, many philosophers working on the free will problem have been attracted to a kind of approach, developed by P.F. Strawson, that justifies belief in free will and moral responsibility by appeal to the essential roles that it plays in our personal and social lives. In this paper I explore some of the limits of this sort of pragmatic approach, arguing that while it may provide a strong justification for treating people as free and responsible in some contexts, especially in our personal relationships, there are reasons to think that this kind of approach is not enough to justify our harshest retributive impulses, especially in contexts like that of a criminal justice system.

Keywords
Free will, moral responsibility, compatibilism, incompatibilism, criminal justice

In recent decades, a kind of pragmatic approach to questions of free will and moral responsibility has gained popularity. The popularity of this approach can, at least in large part, be attributed to transformative work of P.F. Strawson (see Strawson 1974). Strawson, and many who follow in his footsteps, argue that belief in free will—in particular the sense of free will needed to ground moral responsibility and the practices connected to it—is justified by the essential role it plays in our personal and social lives. Like many others, I find this kind of pragmatic approach very appealing—but only to a point. In this paper, I would like to explore some of the limits of this approach. I will argue that the Strawsonian framework can provide strong justification for holding people responsible in some contexts, especially in our personal relationships, but that the pragmatic considerations invoked by Strawsonians are not enough to justify our harshest retributive impulses, especially in contexts like that of a criminal justice system.

I. The Strawsonian Framework

First, I would like to very briefly sketch what I take to be the Strawsonian view.[1] I won't have room to defend it at great length in this paper, but I would at least like to say

1. Of course, people interpret Strawson differently, and take different lessons from his work. What I will sketch here is not meant to be a definitive exposition of Strawson's own view. Rather, I aim to sketch some

a bit about why I think many (including myself) find it to be important and compelling. It will undoubtedly strike many readers as strange, if not entirely misguided, to try to ground the existence of free will and moral responsibility in any kind of pragmatic considerations. Many skeptics argue, plausibly enough, that the fact that the belief in free will and moral responsibility is so central to our personal and social lives tells us nothing about whether that belief is true. It would be similar to arguing, for example, the claim that belief in God is necessary for a meaningful and fulfilling life would, even if true, be no epistemic justification for believing in God (though it might be some pragmatic justification).

In response, I would argue (in line with Strawson, and many others) that claims about freedom and responsibility are fundamentally normative. There is no metaphysical feature of the world we can point at to demonstrate the appropriateness of blame, or to show that a particular agent at a particular time is deserving of praise or gratitude. Claims like these are not existence claims, like claims about the reality or non-reality of God. Rather, they are, at their core, claims about how we ought to regard and treat both others and ourselves, about which kinds of emotions and which kinds of social practices are deserved or fitting or appropriate, and which are not. When making normative claims of this sort, as opposed to simple existence claims, pragmatic considerations regarding the nature and quality of our lives, our relationships, our self-esteem, etc.—claims that are intimately connected with regarding others and ourselves as morally responsible agents—become relevant to the truth of those claims.

To say is *not* to deny the relevance of metaphysical considerations to claims about freedom and responsibility. In my view Strawson and some others are mistaken to conclude that metaphysical considers are completely irrelevant to claims about freedom and moral responsibility. This is because some metaphysical considerations are *included* in the normative standards involved in evaluating the appropriateness of praise and blame. As Gary Watson famously argued, for example, when we learn enough detail about precisely how someone came to be the kind of person they are, even if the person in question is someone truly monstrous (as in his Robert Harris example), our intuitive judgments of freedom and responsibility can be substantially altered (Watson 2004). Drawing on considerations like these, a number of philosophers, working well within the Strawsonian framework, have developed strong arguments for skepticism about free will and moral responsibility.

---

broad lessons that I, and I think many others, draw from taking the sort of approach to the problem of free will and moral responsibility that Strawson did.

In what follows, I will very briefly sketch a few different sorts of skeptical worries that challenge belief in free will and moral responsibility. My aim in this paper is not to evaluate whether or not any of these arguments ultimately succeed. I only wish to show that the following skeptical arguments are at least prima facie plausible, but not (at least given the current state of the dialectic regarding free will and moral responsibility) decisive. The central question I want to consider is how viable a pragmatic justification for belief in free will and moral responsibility (and the practices connected to this belief) is in light of such worries.

II. The Standard Incompatibilist Arguments

A main source of skeptical worry is, of course, the standard arguments for incompatibilism. In recent decades in particular, some powerful new incompatibilist arguments have been developed and much discussed. There is Peter Van Inwagen's familiar Consequence Argument, considered by many to be among the strongest arguments for incompatibilism. It is given various formulations—here is a relatively informal one:

> If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (Van Inwagen 1983, 56)

It is easy to see the force of this argument. The fixity of the past seems beyond question, and it seems plausible to say that event which is a necessary consequence of something unchangeable would itself be unchangeable. This argument motivates what has been termed a 'leeway' condition for free will—that we act freely and responsibly in a given situation only if some other action was possible.

Other standard incompatibilist arguments work to motivate what have been termed 'source' conditions for free will—the idea that we act freely and responsibly only if we are in some deep sense the *ultimate source* of our actions. Versions of this view have been developed and advocated by a number of different philosophers, notably Galen Strawson (Strawson 1994) and Robert Kane (Kane 1996). This intuition has been cleverly defended by the use of manipulation arguments, developed most notably by Derk Pereboom with his famous four-case argument (Pereboom 2001).

Now of course, many incompatibilists who accept arguments like these are not skeptics—they are libertarians who believe that we (at least sometimes) act with fully

free will. Nonetheless, if either these kinds of classical incompatibilist arguments succeed (or any other incompatibilist arguments), that **increases the probability** that we lack the kind of free will that could ground true moral responsibility, that could legitimize praising and blaming people for their actions. If incompatibilism is right, then we have to be able to rule out causal determinism—as well as demonstrate that we have the right kind of indeterministic control for robust leeway and genuine self-creation—before we can know whether or not we have free will.

And indeed, many libertarians explicitly acknowledge that we have little or no epistemic justification for such beliefs. As Richard Double notes in his excellent paper on the 'hard-heartedness' of libertarians, most libertarian thinkers—Roderick Chisholm, Richard Taylor, Peter van Inwagen, Robert Kane etc.—provide very little in the way of any kind of positive evidence that we are the kind of uncaused, self-created entities that satisfy the metaphysically robust conditions for free will and moral responsibility that their accounts demand (Double 2002). As Double discusses, many are quite explicit in admitting that we have no evidence for such claims, notably Immanuel Kant and William James.

III. The Difficulty of Moral Growth

Now I want to briefly discuss a different source of skeptical worry about free will and moral responsibility. The source of worry is based in primarily psychological considerations, rather than philosophical ones.[2] Let us start with a modest philosophical assumption, assumed by almost all who discuss the free will problem. The assumption is that a necessary condition for the kind of freedom that grounds moral responsibility is that we be able to exercise some degree of control over the development of our moral characters over time. The assumption of a capacity for moral self-cultivation is most explicit in a number of libertarian accounts, as described above. But a number of compatibilists, especially in recent years, have articulated the idea that a condition on freedom and responsibility is some sense of self-cultivation.

For example, Al Mele diagnoses a number of variants of Pereboom's "four case argument," saying:

2. For this point, and for much of the discussion in this section, I am heavily indebted to Michael Slote. Slote develops an extensive argument for moral responsibility skepticism along these lines in an unpublished manuscript for a new book on free will, which he was generous enough to share with me.

> In each case in this series, Plum played no role at all in shaping his procedure for weighing reasons (say, through trial and error over the years he has been in the business of deliberating). Unlike normal agents, Plum had no control throughout history as an agent over this important aspect of his deliberative style. (Mele 2005, 78)

In this, Mele suggests that normal agents—agents who are responsible for their actions—shape important aspects of their characters, such as the ways in which they weigh moral reasons in deliberation, over time. This is meant to be a compatibilist reading of "control"—Mele is not assuming any radical contra-causal ability to change one's character of the sort a libertarian might insist on. In Mele's view, what is important is that the development of one's character in influenced and shaped by (does not "bypass") the exercise of one's deliberative capacities over time.

It is easy to see why even compatibilists would be inclined to develop an account of responsibility that requires an ability to shape the development of our own moral characters. The kind of character one has, after all, determines the kinds of actions one commits. And so if one's character is *not* within one's control (even in a minimal compatibilist sense), then it would seem that one's actions would also be (to whatever degree actions are driven by character) outside of one's control. As Michael McKenna puts the point, "what is so important about an agent's having a history that lacks the acquisition of pertinent values through means bypassing her ability to critical assess them is that she thereby *has* a history that afforded her an opportunity to shape her moral personality for herself" (McKenna 2012, 169).

The question is whether this is in fact a psychologically realistic claim about human agency. How flexible are our characters really? To what extent do we really shape and cultivate our own moral personalities over time? And to the extent that we actually can improve our characters over time, to what extent is this really driven by internal processes, or to what extent does it depend on outside help?

One source of pessimism about these questions is research that shows that character traits measured in very young children can have considerable predictive power regarding how they turn out later in life. A famous example of this is the "Stanford marshmallow experiment," a series of delayed gratification studies led by Walter Mischel (Mischel et al 1972). In the studies, young children (ranging from about 3.5 years to 5 years 8 months) were given a choice. They could either take one treat now (like a marshmallow), or they could wait until the researcher came back into the room, in which case they would get two treats. Some children had the self-restraint to delay gratification and wait until the

researcher returned (around 15 minutes), but others could not and would opt for the immediate lesser treat. A series of follow up studies showed that the young children who were better at delaying gratification performed better on a number of measures—they were judged more competent by their parents, they scored higher on SAT exams, performed better on cognitive tests. The original participants even showed more activity in the prefrontal cortex in brain scans conducted in middle life.

Results like this are striking; few would expect that a 4 year old child's capacity to resist the impulse to immediately eat a treat could have such predictive power regarding a person's success in later life. This capacity—now called 'executive function'—is an important part of one's character, and these studies seem to suggest that it is fixed quite early in life. However, this conclusion has recently been challenged. There is now emerging evidence that the capacity for executive function is more malleable than Mischel's initial research suggested, and can be enhanced with the right sort of intervention and training, especially if the intervention is done in early childhood (Zelazo and Carlson 2012). This is certainly good news, but even if right, it doesn't provide much support for the idea that we are in control of the development of our own characters. When it comes to executive function, at least, it seems that making any sort of improvement over the initial capacity we have early in life depends on substantial outside intervention.

And the same may be the case for other, even more clearly morally significant, character traits. Michael Slote discusses the role that empathy plays in moral education (see for example Slote 2010, especially the first chapter), building on the work of the psychologist Martin Hoffman (Hoffman 2001). Hoffman develops the idea that instilling genuinely altruistic, moral motivation and behavior in children requires a process he calls 'inductive discipline,' sometimes simply referred to as induction. Unlike "power-asserting" strategies of moral education (which involve threats and punishments), induction builds from a child's natural initial capacity for empathy. In induction, parents (or other educators) notice when a child has hurt another and then, in a firm but non-threatening way, direct the child's attention to the harm he or she has caused, getting the child to focus on and feel how things must feel for the one that the child has hurt. This leads the child to *feel* the badness of what he or she has done, a painful emotional experience that is a kind of rudimentary guilt. Hoffman argues that if this technique is applied consistently over time, the child will develop an association between these bad feelings and situations in which harm could be (but is not yet) done, without any intervention from parents or other figures, and this will help motivate altruistic and moral behavior.

The key thing to notice about this model is the extent to which the cultivation of moral motivations and moral behavior, via the cultivation of empathy, depends

on parental intervention.[3] This is not *self*-cultivation of moral character; instead, this suggests that the cultivation of moral character depends heavily on others. Without crucial guidance, the early development of one's empathy and one's moral character will be stunted, and later development will be extremely difficult.

Of course, this kind of model is controversial. On some more rationalist-leaning views, empathy is regarded as unnecessary for moral motivation or understanding or growth. I don't have room to say much about that dispute here, but I will note that the claim that empathy is at least a psychologically necessary component of moral understanding and moral motivation for people (even if not logically necessary—maybe some other sorts of possible creatures could grasp and be motivated by morality without it) is at minimum very plausible, and seems well supported by a good amount of psychological evidence. It is well established that individuals who possess little empathy or lack it entirely (in particular associative empathy—the ability and tendency to feel what others feel) have difficulty with both moral understanding (for example, they have trouble drawing a distinction between arbitrary "conventions" and "morality") and moral motivation. Insofar as this model is plausible, our confidence in the idea that we in any substantial sense craft or shape our own moral characters should be lessened.

I think it should be said that even if all of this is right, we still might be able to exercise *some* level of control over our characters. As Neil Levy notes, we may still exercise a kind of indirect control over our characters—we can attempt to engage in long-term projects aimed at altering our characters (Levy 2002). For example, a person with anger management issues might take classes to learn how to meditate in an effort to become calmer and more amiable in his interactions with friends and family. Or a man with a prejudice against a particular ethnicity might embark on a project of studying the history and literature of the group he is biased against to cultivate deeper understanding and empathy with the aim of overcoming his prejudice.[4] Or a woman might buy an app like "HabitRPG" to channel her love of video games into the cultivation of good work habits.[5] And so on.

3. Again, this is a point for which I am indebted to Slote.

4. Slote discusses an example somewhat like this in his unpublished manuscript.

5. HabitRPG is an app allows people to play a sort of video game, in which they gain familiar rewards (gold, experience, levels, etc.) and risk consequences (loss of health, lives, levels) based on successes or failures at pursuing real life goals. For those of us who have cultivated video game addictions, it can be a highly motivating system for cultivating new habits and behaviors.

There are a few things that are important to note about this kind of indirect ability to shape one's own character. As Levy reminds us, one's character is one's way of seeing the world. The impetus to try to make changes to one's way of seeing the world will seldom happen without substantial outside influence. Further, in using techniques like taking classes, or getting therapy, or even using an app, we are relying heavily on assistance from others to shape our characters. Finally, even with a great deal of help from others, long-term efforts to change or improve character are often met with failure, or only partial success. Changing one's character, with or without help, is extremely **hard**.

All of these points strongly suggest that the control we have over our own character is very limited. It might be a substantial enough kind of control to warrant some kinds of praise and blame in some contexts, especially at the level of personal relationships. But it is enough to ground our most extreme negative reactive attitudes, the kind of wrath or hatred that might drive violence? Is it enough to justify a heavily punitive criminal justice system like ours? This is where things start to seem more dubious, or so I will argue.

IV. In (partial) defense of Strawson

At this point I would like to return to the question of the extent to which a pragmatic approach to questions of free will, and in particular moral responsibility, can be justified. As many have argued, our general view of ourselves and of others as morally responsible agents is deeply connected to our relationships with others and our conceptions of ourselves. Freedom and moral responsibility are essential to the possibility of attitudes like love, admiration, and respect, both for others and ourselves. To abandon the concepts of freedom and responsibility is to severely diminish our emotional and moral lives in many ways. As I suggested earlier, given that the normative role played by claims about moral responsibility, pragmatic considerations such as these are essential to deciding whether they are legitimate. Now I would like to say a bit more about what I think about these pragmatic considerations—just what, exactly, is lost if claims about moral responsibility are not legitimate? And what isn't?

To start, let me say a bit about what I think is **not** lost. Some have argued (see for example Peter van Inwagen 1983) that morality itself collapses without free will or moral responsibility, that without praise and blame there can be no legitimate talk of moral obligations, nor even of right and wrong. One way to reach this extreme conclusion is to start with the idea that determinism means that it is impossible for us to do otherwise than we actually do, and then to argue, in Kantian fashion, that this would mean that there can be no such thing as moral obligation (nor of praise and blame). From there

one can argue (as Ishtiyaque Haji does) that there is no such thing as moral rightness or wrongness; "S has a moral obligation to perform [not to perform] A if and only if it is morally wrong for S not to perform [to perform] A" (Haji 1999, 183).

I want to suggest that perhaps the concerns raised by these sorts of arguments are exaggerated. I agree with those who have argued that there are many commonplace examples in which people have moral obligations that they are unable to fulfill. To use an example of Bruce Waller's—if I borrow a large sum of money from a friend, and then hit financial hardship and am unable to repay the loan, it is not as if I am suddenly relieved of my moral obligation to repay. Rather, it seems more natural to simply say that I am now stuck with an obligation that I cannot fulfill. Or consider an example from antiquity.[6] In the Greek tragedy Antigone, the title character finds herself with both an obligation to bury her brother and an obligation to follow the king's law, which prohibits the burial. As Waller notes, "To the Greeks, this seemed an unfortunate situation, but certainly not impossible" (Waller 2011). Haji and some others *do* claim it is impossible, but it is not obvious why this should be so. The claim that we can sometimes have conflicting moral obligations seems to be at least as intuitively plausible as the claim that ought always implies can in every instance. This discussion is a rather quick sketch for the sake of brevity; the minimal point I want to make here is that one can still plausibly maintain belief in moral obligations, and moral rights and wrongs, even if we abandon talk of praise and blame. This point will be important for what I have to say in the next section.

Now I would like to turn to what we might plausibly think **would** be lost if we were to abandon moral responsibility. I think that a number of important moral attitudes would be lost, or at least significantly diminished. For instance, I don't think there can be sincere regret or apology in the absence of moral responsibility (contrary to what some skeptics, like Waller and Pereboom, argue). Waller and Pereboom are right to say that one can *lament* that one is the cause of harm to another in the absence of moral responsibility, or one can lament that one has failed to live up to one's moral obligations. But true **regret** and true **apology** essentially involves taking or accepting responsibility for one's failings. Kathleen Gill puts the point nicely when she argues that an apology without an acceptance of moral responsibility is like saying "I'm sorry" when hearing that a neighbor has leukemia—a mere expression of compassion or sympathy rather than a true apology (Gill 2000). Such expressions of compassion and sympathy are certainly nice, and they definitely have their place, but if all of our apologies were reduced to this,

6. This example is discussed by Bruce Waller (2011), Joseph Margolis (2000), and many others.

then it seems that an essential component of our relationships with others would be missing.

I think this point about regret and apology can be bolstered if we consider a kind of argument common to moral responsibility skeptics. Skeptics commonly argue that there is no morally significant difference between a person in a causally determined universe and a person who is thoroughly manipulated (by say, an evil neuroscientist). This is what Pereboom attempts to show with his four-case argument. And this generally seems to be the view of incompatibilists about moral responsibility and causal determinism. As Waller puts the point, "why should the shaping by fortuitous contingencies not undercut freedom if the same shaping by planned contingencies does?" (Waller 2011, 64). Similar points are frequently made by libertarian incompatibilists. For an example of this, see Robert Kane's discussion of B.F Skinner's "Walden Two" story (Kane 1996, Chapter 2).

So let's grant for the sake of argument that incompatibilists like Waller and Pereboom and Kane are right, that there is no morally significant difference between a causally determined agent and one who has been manipulated by an outside agent. And then let us ask—to what extent could a manipulated agent truly regret her actions? To make the question more concrete, let's consider a specific example. Imagine a woman named Riley who is being completely manipulated and controlled by a wicked neuroscientist who has planted a device in her head. One day Riley sees a child drowning. She has an impulse to save the child, but that impulse is quickly erased by the neuroscientist, who replaces it with an irresistible desire to turn and walk away instead.

Now suppose that after walking away from the beach and knowingly allows the child to drown, Riley then later learns that her actions had been directly programmed and controlled by a nefarious neuroscientist. It seems that Riley would be right to believe that she was not blameworthy for letting the child drown. Could she at the same time sincerely regret her action? It seems clear to me that she could not. Riley might be extremely sad that the child had drowned, and she might lament the fact that she had been used as a tool by the neuroscientist to bring about the child's death. But insofar as she *truly* regards the neuroscientist's manipulation as completely undercutting her moral responsibility, it is hard to see how she could genuinely regret the action. If this is right—and if incompatibilists are right that there is no morally significant difference between manipulation and ordinary causal determinism—then it is also hard to see how a moral responsibility skeptic can say that it would *ever* be appropriate to experience true regret. The only way that I can see for such a skeptic to avoid this conclusion in the ordinary deterministic case would be to admit that there is a substantial moral difference

between manipulation cases and ordinary causal determinism—but this admission would undermine one of the major incompatibilist strategies for defending their position.

The same goes for positive corollaries of reactive moral attitudes like regret and sorrow, for instance attitudes like appreciation and gratitude. The reason why positive moral attitudes like gratitude are threatened by the demise of moral responsibility is very similar to the reason why regret and sorrow are threatened. The reason is that a central component of such attitudes is the belief in the sort of freedom required for moral responsibility—the belief that the person to whom you grateful is an apt target for praise and blame for his actions. As Galen Strawson writes, "It seems that we very much want people to be proper objects of gratitude, for example. And they cannot be proper objects of gratitude unless they can be truly responsible for what they do" (Strawson 1986, 308). Lucy Allais expresses the point similarly, saying "feeling gratitude towards someone with respect to an action involves seeing the action as flowing from her free choice" (Allais 2008, 179).

Even Pereboom concedes this point to an extent, saying, "Gratitude might well require the supposition that the person to whom one is grateful is morally responsible for an other-regarding act, and therefore hard incompatibilism might well undermine gratitude" (Pereboom 2001, 201). Pereboom says that we can still have a sense of "thankfulness" in the absence of true gratitude (a kind of thankfulness that Waller seems to equate with true gratitude), suggesting "one can also be thankful to a pet or a small child for some favor, even if one does not believe that he is morally responsible. Perhaps one can even be thankful for the sun or the rain even if one does not believe that these elements are backed by morally responsible agency" (Pereboom 2001, 201).

In my view, examples like these highlight just how far removed the attitude of "thankfulness" that we might have towards those we regard as lacking moral responsibility is from genuine gratitude. Certainly we can, as Pereboom suggests, experience joy and thankfulness when someone (or something) who lacks moral responsibility does something nice for us. But I think we want something much deeper than this out of our relationships. If the gratitude and appreciation that we can have for our dearest loved ones is diminished to the level of the kinds of emotional reactions that I can have to pets or even blind forces of nature, then it seems that something very substantial about our personal relationships has been lost.

I would also like to say a little bit about the connection between love and freedom and moral responsibility. The idea that genuine freedom and moral responsibility might be essential for love has been expressed, to different degrees, by a number of philosophers. P.F. Strawson himself claims that the range of emotions we can experience without the

moral reactive attitudes "cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other" (Strawson 1974, 10).

Some ways of arguing for this claim are misguided. For example, some, like Robert Kane, argue that the most valuable sort of love is that love which is freely chosen. And W.S. Anglin argues that love that is necessitated (whether by manipulation or coercion or the causal structure of the world) is not authentic (Anglin 1999). This idea may have some initial intuitive appeal, but it immediately runs into obvious objections. Pereboom mentions the example of familial love, such as the love between a parent and child (Pereboom 2001). It seems completely implausible to suggest that, for example, there is any exercise of will (free or otherwise) involved in the instantaneous bond of love that forms between a mother and a newborn child. In fact, it would seem inappropriate for such a bond to have to be mediated by any effort of will on the part of the mother, free or otherwise. If the mother had to actively will herself to love her new child, we would take it as a sign that something was awry. In this instance, completely unwilled, unfree love seems to be the ideal. The same can be argued for romantic love. As Nomy Arpaly reminds us, there is a sense in which we find it romantic to say "it had to be you"—to express the fact that there is no possible way I could fail to love you (Arpaly 2006).

So I think the claim that moral responsibility is necessary for love because love must be freely chosen is mistaken. Still, moral responsibility does, in my view, play an important role in grounding our loving relationships. Consider, for example, the essential role played by the emotions discussed above—gratitude, regret, sorrow, and related attitudes like forgiveness—in loving relationships. Insofar as these are an essential component of fully deep, authentic loving relationships between adults, loving relationships are deeply diminished in the absence of moral responsibility. To see this, just imagine a relationship with a person who is regularly manipulated (pick your favorite evil neuroscientist manipulator story) in ways that rob her of responsibility for her actions. She does kind things for you sometimes, other times she is thoughtless or hurtful, but in all of her interactions she is thoroughly manipulated in ways that rob her of responsibility, that make it impossible to feel deep gratitude towards her for her kindness, or for her to truly regret her bad behavior or take responsibility for it, etc. You might feel some strong affection for her, even a kind of love, but it seems to be it would be substantially diminished in comparison to that which we feel for those who we believe to be the apt targets of the reactive attitudes that comprise moral responsibility.

In this section I hope to have fairly characterized the kind of case that I think can be made in defense of the importance of moral responsibility drawing on the kinds of

considerations that Strawson first drew our attention to. Ultimately I think it is a strong case. In the next section, I want to a bit more about how far I think this case can be extended, and offer some suggestions about where its limits may lie.

V. The Limits of the Reactive Attitudes

Now that I have said a bit about the ways in which I think some pragmatic considerations—in particular connected to the nature and quality of our relationships with others—can justify and ground moral responsibility, I want to explore in a bit more detail some of the limits of this justification.

First, I want to suggest that the strength of this kind of justification varies according to the context in which someone is being held responsible for his or her actions. I think this kind of justification is strongest in the context of every day life, in our ordinary kinds of interactions with people. As I argued in the previous section, the best defenses of the moral responsibility system are connected to the role it plays in our lives and our relationships. It is essential for attitudes like gratitude and regret, sorrow and pride, attitudes that are essential for our loving relationships, and also to our regard for ourselves. In that context, it makes sense to say that people deserve the ordinary kinds of reactive attitudes and treatments (positive or negative) that their (positive or negative) treatment of others invites.

And I think that in this context, the skeptical worries that I have raised so far are at their weakest. Consider for example the standard incompatibilist arguments I sketched earlier, the ones that suggest source or leeway requirements for moral responsibility. As I argued in the last section, it seems implausible to say that love hinges on any claims about people being the ultimate source of their love, or of having any choice in the matter at all. We care that people we love be autonomous in some sense—it would seem difficult to feel genuine love, or to have a full range of moral reactive attitudes, for a thoroughly manipulated agent—but it strains credulity to suggest that love requires contra-causal freedom of the sort that incompatibilists insist on.

Likewise, I don't think that the skeptical worries raised by the difficulty of self-orchestrated moral change and moral growth pose a very strong a threat to moral responsibility in ordinary detail contexts and in our personal relationships. On the contrary, these considerations may even help in some ways to support the importance of holding people (both ourselves and others) morally responsible for their actions. One of the points I emphasized in that section is that moral change and moral growth often requires substantial input and help from others. Many defenders of moral responsibility

(especially compatibilists) have appealed to the communicative role of our moral responsibility practices. One of the ways we communicate our moral expectations to others is in our emotional expressions—in our approval or disgust or shame or gratitude or anger, etc. And through these communications, via the moral reactive attitudes that comprise our moral responsibility practices, we can help one another to grow morally (for a detailed discussion of evidence supporting this, see Shaun Nichols 2007).

In short, given that the reactive attitudes are constitutive of so much that is essential to our relationships and our self-regard, and given they have an important role to play in how we grow and develop as moral agents, skeptical arguments carry less force in this context. The Strawsonian picture is most compelling here.

I think things are a bit different, however, when we shift to a context like criminal justice. Here our judgments of responsibility and praise and blame have much more serious consequences. When talking about criminal justice and criminal punishment, the stakes are very high. When we incarcerate criminals, we deprive them of liberty and subject them to conditions that are severe impediments to living a life of any kind of quality. For severe crimes we sometimes even deprive criminals of their lives. And the justification for this sort of practice is closely tied with moral responsibility. As Stephen Morse puts it, "both the criminal and the medical-psychological systems of behavior control require a justification in addition to public safety—desert for wrongdoing or non responsibility (based on disease)—to justify the extraordinary liberty infringements that these systems impose" (Morse 2013, 29).

There are two important points to be made here. The first is that when the stakes are this high, the epistemic standards should be raised raised. If the justification for a criminal justice system that deprives people of liberty is going to be grounded in moral responsibility and desert, then the justification for believing that criminals in a particular instance are responsible in the sense that could ground desert must be very strong. Pereboom expresses this point as follows: "As I argued in the context of criminal punishment, if one aims to harm another, then one's justification must meet a high epistemic standard. If it is significantly probable that one's justification for the harmful behavior is unsound, then it is best that one refrain from engaging in it" (Pereboom 2014, 318). What Pereboom expresses here seems right. Even if the kinds of skeptical arguments I've discussed in this paper fall short of being decisive in the context of a criminal justice system, insofar as they raise significant doubts and lower our level of credence in our convictions about the moral responsibility of criminals, they do provide a strong reason to exercise restraint in criminal punishment. Even if we think the odds of moral responsibility skepticism being the correct view is fairly small, it still may be

reasonable to judge that the risk that we may cause great harm to people who do not deserve it (on the small chance that moral responsibility skepticism is correct) is morally significant enough to revise our criminal justice system and treat criminals somewhat more as we should if skepticism were true.

The second point is that the skeptical arguments carry much more force when considered in the context of criminal justice. As I argued before, the claim that ultimate sourcehood or leeway conditions, in the incompatibilist sense, are necessary to ground authentic love or the moral reactive attitudes in our personal relationships is not very compelling. But I think that these arguments are much more compelling when we are talking about the kind of responsibility involved in justifying deprivation of life or liberty or other seriously harmful punishments. An incompatibilist requirement like the requirement that one be the ultimate source of his or her character makes much more sense when trying to argue that one deserves something as severe as capital punishment (for example) because of their wrongdoing. Similarly, the worry that we might have a very limited capacity to shape our own moral characters without input from others, a worry that raises substantial problems of moral luck, is most pressing when we are talking about inflicting serious harm on people for the crimes that their characters drive them to commit. There is a reason that many skeptics (like Waller and Pereboom) focus heavily on questions of criminal justice and social justice when advancing incompatibilist arguments like these—because it is in these contexts that the arguments carry the greatest intuitive force.

VI. What Sort of Criminal Justice System Should We Have?

The question that remains now is what should our criminal justice system be like? What are the costs of altering or giving up (at least some) of our traditional ideas of moral responsibility and blameworthiness in the context of criminal justice? I have argued that abandoning the idea of moral responsibility in our daily lives diminishes our relationships with others and our self-esteem. But would anything comparable happen if we were to modify our criminal justice system, focusing less on the suffering that criminals might or might not deserve, and instead—as a skeptic would prescribe—more on forward looking considerations (see Pereboom 2014) like rehabilitation and crime prevention? In my view there is no strong reason to think this.

On the contrary, there are several good reasons reason to worry about a justice system that places too much emphasis on retributivism. For starters, there is the worry that a justice system that places too much emphasis on retributivism will be limited in

the extent to which it engages in investigating and learning about the causes of people's actions. This may not be a limitation that exists as a matter of logical necessity, but nonetheless, it does seem to be a common feature of highly retributivist societies with justice systems that put the main focus on making sure that criminals 'get what they deserve.' The basic worry is that the more we as a society are inclined to judge, the less we are inclined to try to understand. But when it comes to setting social policies, it is understanding—of psychology, sociology, economics, the effects of punitive and rehabilitative and other social policies—that we need. Waller offers a striking example of this extreme kind of retributivist attitude: "As the British Prime Minister, John Major called for harsher criminal justice measures, especially against juveniles: 'Society needs to condemn a little more and understand a little less'" (Waller 2011, 283).

There is also considerable evidence that justice systems that focus more heavily on retributivism—on harsh punishments, on making sure that criminals 'get what they deserve'—produce worse outcomes. Optimistic free will skeptics have highlighted much of this data. For example, the American justice system is well known to be one of the harshest in the world, and it has been argued that this is closely connected with our sense of 'rugged individualism' and belief in absolute individual responsibility for our actions (for example see Waller 2011, 282–287). Since 2002 the U.S. has incarcerated a greater percentage of its population than any other nation in the world—about 500 prisoners per 10,000 people, or 1.6 million prisoners total, in 2010 (see Guerino, Harrison, and Sabol 2012). The numbers get even higher if we include jails as well as prisons. We are one of few nations to retain the death penalty, we use 'life imprisonment' for a wide range of crimes in comparison to most other nations, and have continually expanded minimum sentencing laws and the use of 'three strike' laws. And yet there is little evidence that our continually increasing 'toughness' on crime has produced a significant deterrent effect. A major review of studies of the deterrence effect of harsh sentences found "…the studies reviewed do not provide a basis provide a basis for inferring that increasing the severity of sentences generally is capable of enhancing deterrent effects" (von Hirsch, Bottoms, Burney, Wikstrom 1999). These facts are well known, and yet there is little political will to soften or revise our sentencing guidelines—arguably because we are so driven by a need for retributive justice.

The question that remains now is what sort of criminal justice system should we have? In light of worries like those raised above, in addition to the skeptical arguments we have considered, is there *any* role for retributive considerations? I want to suggest that perhaps there still is. First, I would like to acknowledge Morse's point that if we are ever going to deprive people of liberty, we must have good moral justification. Skeptical

arguments are significant enough that we should be less punitive than we often are, less driven by the desire for revenge. We should err on the side of compassion and mercy when we are able, and focus more heavily on outcomes rather than deserts. But nonetheless, we can at the same time consistently say that criminals *do* deserve *some* level of approbation and punishment, and this sense of desert can be motivated by the practical considerations outlined above. And a grounding in some notion basic desert is important if are to avoid the moral problems that arise from a criminal justice system grounded purely in consequentialist considerations.[7]

And as a further suggestion, I would just like to briefly mention one natural way to incorporate the kind of Strawsonian view I have defended in the context of personal relationships into a criminal justice system. This way can be found in the idea of restorative justice. Restorative justice is an approach to justice that focuses on the circumstances and needs of the victim, and the victim's relationship to the transgressor. Offenders are encouraged to take responsibility for their actions, and enter into a dialog with the victim to apologize and (depending on the nature and circumstances of the crime) to offer some way of making amends. Victims play an active role in determining what punishment the offender will receive. This approach to criminal justice resembles the Strawsonian view of responsibility, as grounded in our relationships with others, which I have defended in this paper. It avoid the abstract and extreme notion of desert that infests our criminal justice system as it exists, a notion which can lead to extreme sentencing, and which (as I have argued) is more vulnerable to skeptical worries. It recognizes that what one deserves for committing a crime in part consists in the effects on the victim and the needs of the victim, and can be shaped by one's relationship to the victim—even if that relationship is formed after the fact in the restorative justice process. Of course, much more needs to be said to defend and refine this approach to justice, and that would take me beyond the scope of this paper (see Sommers 2013 for an excellent exploration and defense of this kind of approach). But I do think that this approach at least holds promise—and it would be a way to develop our justice system in accordance with the kind of moral responsibility that I have argued can be well justified by pragmatic considerations.

7. I don't have time to explore these problems in detail here, but to give just one example, consider Saul Smilansky's argument that without *any* kind of moral desert, we would be morally required to make the lives of criminals as comfortable and enjoyable as possible—to give them 'funishment' instead of punishment (Smilansky 2011).

VII. Conclusion

In this paper I have presented arguments that suggest that there are limits to the kind of moral responsibility that can be justified by Strawsonian-style pragmatic compatibilist considerations. Some of these arguments have been admittedly somewhat briefly sketched. Still, I hope to have made a plausible case that while pragmatic considerations can arguably provide strong grounds for moral responsibility in the context of our daily lives (strong enough to resist the major skeptical worries), this strategy is much weaker when used to try to ground a harshly retributive criminal justice system—in particular one that resembles what exists in America today (as well as several other nations). In sum, the pragmatic justification I have been considering in this paper, the sort of justification that seems to provide strong grounds for regarding both others and ourselves as apt targets for moral reactive attitudes in the personal domain, doesn't seem adequate to justify the abstract and extreme concept of desert that seems to operate in the domain of justice. If we want to find a role for retribution and responsibility in justice, then I suggest that we need to reform our justice systems to more closely model the features of our personal relationships that provide a solid footing for the reactive attitudes in the first place.

References

Allais, Lucy. 2008. "Dissolving Reactive Attitudes: Forgiving and Understanding." *South African Journal of Philosophy* 27 (3): 179–201.

Arpaly, Nomy. 2006. *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton, NJ: Princeton University Press.

Double, Richard. 2002. "The Moral Hardness of Libertarians." *Philo: A Journal of Philosophy* 5 (2): 226–234.

Gill, Kathleen. 2000. "The Moral Functions of an Apology." *The Philosophical Forum* 31 (1): 11–27.

Guerino, Paul, Paige M. Harrison, and William J. Sabol. 2012. "Prisoners in 2010." *Bureau of Justice Statistics*.

Haji, Ishtiyaque. 1999. "Moral Anchors and Control." *Canadian Journal of Philosophy* 29 (2): 175–203.

Hoffman, Martin. 2001. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge: Cambridge University Press.

Kane, Robert. 1996. *The Significance of Free Will*. New York: Oxford University Press.

Levy, Neil. 2002. "Are We Responsible For Our Characters?" *Ethic@* 1 (2): 115–132.

Margolis, Joseph. 2000. "Excerpts from Ishtiyaque Haji's Discussion with Members of the Audience." *The Journal of Ethics* 4 (4): 368–381.

McKenna, Michael. 2012. "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism." *J Ethics* 16 (2): 145–174.

Mele, Al. 2005. "A Critique of Pereboom's 'four-case Argument' for Incompatibilism." *Analysis* 65 (285): 75–80.

Morse, Stephen. 2013. "Common Criminal Law Compatibilism." In *Neuroscience and Legal Responsibility*, edited by Nicole Vincent, 27–52. New York: Oxford University Press.

Mischel, Walter, Ebbe B. Ebbesen, and Antonette Raskoff Zeiss. 1972. "Cognitive and attentional mechanisms in delay of gratification." *Journal of Personality and Social Psychology* 21 (2): 204–218.

Nichols, Shaun. 2007. "After incompatibilism: A naturalistic defense of the reactive attitudes." *Philosophical Perspectives* 21 (1): 405–428.

Pereboom, Derk. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

Slote, Michael. 2010. *Moral Sentimentalism*. New York: Oxford University Press.

Sommers, Tamler. 2013. "Partial Desert." In *Oxford Studies in Agency and Responsibility, Volume 1*, edited by David Shoemaker, 246–262. Oxford: Oxford University Press.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon Press.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 (1–2): 5–24.

Strawson, P. F. 1974. *Freedom and Resentment, and Other Essays*. London: Methuen.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Von Hirsch, Andrew, Anthony Bottoms, Elizabeth Burney, and P-O Wikstrom. 1999. *Criminal Deterrence and Sentence Severity: An Analysis of Recent Research*. Oxford: Hart Publishing.

Waller, Bruce N. 2011. *Against Moral Responsibility*. Cambridge: MIT Press.

Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. Oxford: Clarendon Press.

Zelazo, Philip David, and Stephanie M. Carlson. 2012. "Hot and Cool Executive Function in Childhood and Adolescence: Development and Plasticity." *Child Development Perspectives* 6 (4): 354–360.

Journal of Cognition and Neuroethics

A Kantian Defense of Libertarian Blame

John Lemos
Coe College

Biography

John Lemos is the Joseph McCabe Professor of Philosophy at Coe College in Cedar Rapids, IA. He is the author of two books, *Commonsense Darwinism* (Open Court Press, 2008) and *Freedom, Responsibility, and Determinism* (Hackett Publishing, 2013). He has also published numerous articles in a variety of journals, such as *Dialectica*, *Metaphilosophy*, *Philosophia*, and *The Southern Journal of Philosophy*.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Lemos, John. 2015. "A Kantian Defense of Libertarian Blame." *Journal of Cognition and Neuroethics* 3 (1): 251–263.

A Kantian Defense of Libertarian Blame

John Lemos

Abstract
Libertarianism is the view that free will exists and it is incompatible with determinism. As such, libertarians believe that at least some of our free willed acts must be undetermined. Many contemporary libertarians admit that there is not adequate epistemic justification for the view, yet they endorse the view and the practices of praise/blame and reward/punishment which they ground on the presumption of libertarian free will. This article considers a moral objection to this aspect of libertarianism and responds to it with a kind of Kantian pragmatic defense.

Keywords
Free will, Responsibility, Libertarianism, Kant

Determinism is the view that at any time the universe has exactly one physically possible future. Libertarianism is the view that free will exists and it is incompatible with determinism. As such, libertarians believe that at least some of our free willed acts must be undetermined and, thus, that determinism is false. Furthermore, most libertarians believe there is no adequate epistemic justification for belief in the existence of libertarian free will, and most of these same libertarians believe that we should hold people accountable for their actions—blaming them and punishing them when they act wrongly of their own free will.[1]

It could be argued (indeed, some philosophers have argued) that libertarians are acting immorally when they hold people accountable for their actions, blaming and punishing them, while believing there is no adequate epistemic justification for belief in free will. The argument runs as follows:

1) Libertarians believe that we should hold persons morally responsible.

1. William James is a key historical figure who accepted a libertarian view and who believed there is no sufficient epistemic justification for the view. Immanuel Kant was also a libertarian about free will and he believed that while there was no theoretical reason to believe in free will there were, nonetheless, good practical reasons to believe in it. Some recent and contemporary figures who believe in libertarian free will, but who also believe there is no sufficient epistemic justification for such belief are: Roderick Chisholm (1976); Richard Taylor (1966); Peter van Inwagen (1983); William Rowe (1995); Timothy O'Connor (1995a); Robert Kane (1996); and Mark Balaguer (1999). The libertarian views of the latter three thinkers are developed further in: O'Connor (2000); Kane (2007; 2011); and Balaguer (2010).

2) Libertarians believe that we should hold persons morally responsible only if they exercise libertarian free will.

3) Most libertarians believe that we have scant epistemic justification that persons have libertarian free will.

So, 4) most libertarians believe we have scant epistemic justification for believing that persons meet one of the necessary conditions for being morally responsible, while still believing we should hold persons morally responsible.

5) Sympathetic or morally conscientious persons do not hold people morally responsible for their actions unless they have epistemic justification for doing so. To do this is to be hard-hearted.

So, 6) most libertarians are not sympathetic and morally conscientious, i.e. they are hardhearted.[2]

In this essay I will defend libertarianism against such moral criticism. In particular, I will argue that even if libertarians believe there is not adequate epistemic justification for belief in libertarian free will, they can still, nonetheless, be morally justified in holding people morally accountable for their actions. In defending this position, I will argue along basically Kantian lines.

The Kantian Response

Kant's principle of ends states that one should always treat humanity, whether in one's own person or in that of others, as an end and never as mere means. To treat a person as an end is to show respect for the person's autonomy; it is to show respect for that person's ability to make choices for himself and to act in accord with them. To treat someone as mere means is to disrespect this capacity of persons. Rape, murder, theft, slavery all involve treating others without regard for their own choices. The person raped did not choose to have sexual relations with the rapist—the sexual relations are forced upon him/her. The person robbed did not choose to give up his/her property—it is taken

2. See Double (2002) for a recent defense of this line of argument. Derk Pereboom makes a similar point in *Living Without Free Will* (2001), 198–199.

against his/her will. According to the principle of ends, these acts are wrong simply because they involve treating persons as mere means and not as ends in themselves who possess a capacity of choice that deserves respect. According to this principle, rape, murder, and theft are not wrong due to their bad consequences. Indeed, one of the cardinal virtues of this Kantian principle is that it captures the widespread intuition that acts like murder and theft can be wrong even when the consequences of these actions are good on the whole.

Now, as noted, what makes persons deserving of respect is their capacity for choice and their ability to live their lives in accordance with their choices. Choice can be understood along either hard determinist, hard incompatibilist, compatibilist, or libertarian lines. On all of these views, choice is to be understood as the end result of deliberation. There is no doubt that all human beings do frequently engage in deliberations about what they shall do and in doing so they assume that what they will do is up to them. Now on the hard determinist view of things, our choices are never freely performed and we are never responsible for them because what we choose is just a necessary consequence of prior factors which were in turn necessitated by even earlier events and so on going back in time. Hard incompatibilists agree that we never engage in free choice and that we are never responsible for our choices. They believe that all of our actions are either determined or, perhaps, some of them are undetermined, but either way we do not make free choices and we are not responsible for them. If our choices are determined, then they are not free for the reasons indicated by the hard determinists. Furthermore, if they are undetermined, then they are random occurrences, meaning that we lack the kind of control over them for them to be products of free will.

On the libertarian view, choices can be freely made and we can be morally responsible for them. For this to be the case at least some of them must be causally undetermined. The libertarian does not have to view all free choices as undetermined. As Robert Kane has noted, the libertarian can view determined choices as free in a derivative sense if they are the consequence of a character formed by prior undetermined free choices.[3] Compatibilists also believe choices can be free and that we can be morally responsible for them. However, unlike the libertarians, they believe that even if all events, including all of our choices, are determined, then we can still make free choices and be responsible for them.

For the most part libertarians do not think compatibilist accounts of freedom and responsibility make sense. Hence, the famous quips of William James and Immanuel

3. See, for instance, Kane 1996, 2007, and 2011.

Kant, two historically famous defenders of libertarianism; James called compatibilism a "quagmire of evasion" (1884, 149) and Kant called it a "wretched subterfuge (1788, 95–96)." Further, there are some pretty good reasons to think compatibilism is deeply flawed—consider Peter Van Inwagen's consequence argument (1983) or Derk Pereboom's four case argument (2001) or Robert Kane's argument from ultimate responsibility (1996).

A libertarian may reasonably come to believe that there is no plausible compatibilist account of freedom and moral responsibility. If so, he will be led to think there is either libertarian free will or no free will and no moral responsibility at all. This point is very significant in developing a reply to the charge that libertarians are hard-hearted. For when choice is viewed on the hard determinist or hard incompatibilist models it is not perceived as free choice for which the agent is responsible. On these models no human beings ever make free choices for which they are also morally responsible. If choice is understood in these terms, it is hard to see how the human capacity for choice gives us the special dignity and worth that entails we should always be treated as ends and never as mere means. Indeed, for Kant the capacity for free choice was the grounds for thinking of human beings as autonomous beings deserving of respect. Thus, in order to account for this autonomy he was led to conceive of humans as possessing a transcendental (noumenal) self that stood outside the realm of deterministic causal law.

While I have no interest here in defending the notion of a Kantian transcendental (noumenal) self, I think it is correct to believe that we cannot make sense of the Kantian principle of ends unless human choices are perceived as free choices for which we are morally responsible. Furthermore, assuming libertarians are correct in their rejection of compatibilist models, it follows then that we can only make sense of the principle of ends on the assumption that human beings have libertarian free will.

What I am suggesting is that a libertarian who finds compatibilism implausible and who accepts the nonconsequentialist principle of ends on independent moral grounds may be rationally led to posit the existence of libertarian free will and moral responsibility without epistemic support, because this is the only coherent way to make sense of what he knows to be true in the realm of morality.

One might come to believe that the principle of ends is true for independent moral reasons. For instance, one might for independent moral reasons come to believe that the Kantian principle of universalizability is true and then deduce the principle of ends from it. That is, one might consider treating persons as mere means and apply the principle of universalizability in assessing the rightness of acting in such a way. In doing so, he might see that one cannot reasonably will that everyone treat persons as mere means and,

then, conclude that we should always treat persons as ends in themselves. Or, one might notice that many problems with utilitarianism involve instances of unjustified treatment of persons as tools in the pursuit of the greater good, and then one might be led to see that each of these is a case involving the violation of a general rule that persons should always be treated as ends.

Now, having come to accept the principle of ends on independent moral grounds, one might then come to the realization that rational acceptance of it requires that we believe humans have free will. Further, if one has good reason to believe compatibilist accounts of free will are implausible but that there are plausible libertarian accounts, then one might rationally be led to posit the existence of libertarian free will and assume that people are morally responsible. That is, one might rationally assume this without epistemic justification, as this is the only way to make sense of the principle of ends, which one has already come to accept for independent moral reasons.

Epistemic Justification, Kant, and the Nature of My Argument

Before going forward, I would note that some readers may find it odd that I suggest Kant believed there was no epistemic justification for belief in libertarian free will. It might be thought that, according to Kant, we have good evidence for the truth of the categorical imperative and because of this we have good evidence of the existence of libertarian free will. Thus, from the Kantian perspective there is epistemic justification for belief in libertarian free will.

In responding, I would note that Richard Double, a leading proponent of the argument that libertarians are hardhearted, regards Kant as holding that there is no epistemic justification for belief in libertarian free will. Double states:

Immanuel Kant proclaims that we can have no epistemic justification for believing that persons make libertarian choices, but recommends that we postulate on faith alone the existence of trans-empirical selves "in" a noumenal world who (that?) make such choices (Double 2002, 227).

I am following in Double's footsteps by interpreting Kant in this way, and I think there is rational warrant for doing so.

Kant distinguishes between theoretical reason and practical reason. Theoretical reason aims at revealing what is true in the realms of mathematics, empirical science, and pure metaphysics. From the perspective of theoretical reason there is no epistemic justification for belief in libertarian free will. In contrast, practical reason aims at determining what it is right to do. And Kant held that a belief in libertarian free will

was necessary to making sense of our moral understanding. Thus, he was led to posit the existence of libertarian free will, which he believed we lacked adequate evidence for in the realm of theoretical reason. It is in this sense that I, like Kant, want to argue that we may be rationally warranted in believing in libertarian free will without epistemic justification.

Does rational acceptance of the principle of ends require belief in free will?

Is it true that we can only make sense of the principle of ends on the assumption that free will exists? It might be argued that this is a controversial claim and without a good argument for it my critique might rightly be regarded as question begging. Perhaps the reason why people should be treated as ends is that they have the rational capacities that make them able to understand the world and the consequences of their actions and they are able to deliberate and make choices in accordance with such knowledge. It could be argued that even if such choices are not freely made by persons we should still respect the capacities people have to make such choices by treating them as ends.

Such a response would be misguided. To see this consider that someone, call him "the Puppetmaster," has the power and knowledge to take every young child that is not yet of the age to reason and deliberate and impose a set of beliefs and values and reasoning skills upon them such that they would then deliberate, choose, and act in accord with these. Further, imagine that the Puppetmaster is kind and wise and that he endows every child with good values, true beliefs, and sound reasoning abilities. Consequently, when the children begin to think and reason and choose they are always led to make the right decisions. Finally, imagine that once the Puppetmaster has given these children their beliefs, values, and reasoning abilities he does not interfere with them in later life; rather, he lets them think, reason, and choose in accord with the mental programming he has provided for them.

Now, would the Puppetmaster have violated the principle of ends in doing this to every child? It seems fairly obvious that he would have. However, if you think we can make sense of the principle of ends without believing in free will but that we can make sense of it just in terms of the human capacity for choice, then you will have a hard time explaining how the Puppetmaster has violated this principle. For, according to the example the Puppetmaster has in no way intruded upon the ability of persons to make choices. He has simply given them the beliefs, values, and reasoning abilities that will dictate how they will deliberate and choose in later life. Had he not done this, then on the deterministic model genetics and environment would have provided the mental

programming; and done a worse job of it, I might add, since the Puppetmaster provides programming that always leads to right action.

In contrast, if we understand the principle of ends as involving a belief in free will, then we can make sense of the wrongness of the Puppetmaster's action. That is, if we think that persons are to be treated as ends because of their capacity for ***free*** choices, then the Puppetmaster has clearly violated the principle of ends. The Puppetmaster makes everyone such that they choose in accord with the mental programming he has provided for them. While he does not interfere with their capacity to choose, he does interfere with the freedom of their choices in the sense that the agents subjected to his programming do not have the freedom to shape their own beliefs, values, and decision-making style.

It might be objected that I am setting the requirements of human freedom at ridiculous heights—that I am assuming human freedom involves a capacity to create one's own character and that this makes the standard of human freedom unattainable since we must all start with the given of genetic and environmental input. However, such a retort is misguided. I'm only advocating that the kind of freedom needed to make sense of the principle of ends includes the ability to have a role in shaping one's own beliefs, values, and decision-making strategies. This does not require an ability to create oneself ***ex nihilo***. Of course, we have to start with what is given to us from the lottery of genetics and environment, but from there we must have the freedom to critically evaluate our inherited system of beliefs and values and to accept or reject what we are initially given. Compatibilists think we can get the requisite kind of freedom to do this on a deterministic model, but most libertarians don't think this will suffice. Nonetheless, whether we conceive of this freedom in compatibilist or libertarian terms, my point is that without this limited freedom to self-create (not to self-create ***ex nihilo***) which the Puppetmaster would deny us, then we don't have the freedom required by the concept of personhood which is invoked in the principle of ends. Thus, if one does not believe we have such freedom, then one will not be able to make sense of the fact that were the Puppetmaster to do this to every child he would be violating the principle of ends.

Does my appeal to the principle of ends involve me in a contradiction?

Another objection to my argument might note that when we hold people responsible for wrongdoing and blame and punish them while believing that we have insufficient epistemic justification for belief in their guilt, then we treat them as mere means, violating the principle of ends. Thus, if one concedes that there is no

epistemic justification for belief in libertarian free will and one believes libertarian free will is necessary for moral responsibility and one still holds people morally responsible for wrongdoing, then one violates the principle of ends. In this way, my appeal to the principle of ends in support of holding people responsible without epistemic justification involves me in a contradiction.

In response to this, I suggest that in the arena of practical reason—where, among other things, we assess the moral value of our own actions and those of others and we try to determine what is right to do—if we are to rationally employ the principle of ends, then we must assume there is free will. Further, the assumption that there is free will legitimizes attributions of moral responsibility and blaming and praising and punishing and rewarding. If in the sphere of practical reason I am led to wonder what it is right and wrong to do and these quandaries lead me to adopt the principle of ends on the basis of good reasons, then I am warranted in my acceptance of this principle and acting in accord with it even though it commits me to a belief in free will that I cannot epistemically justify in the domain of theoretical reason. If I am right about this, then I can be warranted in my attributions of moral responsibility and in my praising and blaming and rewarding and punishing, and doing so involves me in no contradiction.

Here it might be said that there surely is a contradiction, because we justly say that a person violates the principle of ends when he blames and punishes another person while he knowingly lacks sufficient evidence of his guilt. Thus, if we say that it's acceptable to blame and punish when we know there is insufficient epistemic justification for belief in the existence of free will, then there is a contradiction. For if the principle of ends is violated in the former case, then it must be violated in the latter case.

However, such an argument is grounded on confusion. The reason we take care not to blame and punish persons based on insufficient evidence of their guilt is because we already take them to be ends in themselves with the power of free choice that makes them deserving of the respect which is commanded by the principle of ends. If there's a rattlesnake in the road where my children are playing, then I may kill it or, at least, forcibly remove it from the road to protect them. If there's a suspicious looking person walking in the road where my children are playing, I'm not entitled to kill him nor forcibly remove him from the road. And why not? Because as a human being the principle of ends applies to him and I should not bring harm upon him unless he does through his own free will commit certain acts which merit a response that may be harmful to him.

If the principle of ends is the central element of my moral outlook, then I will hold to the principle that people should be regarded as innocent until there is sufficient evidence of their guilt. But a proper understanding of the principle of ends is grounded on the

presumption of free will. Consequently, if rational moral considerations lead me to the adoption of the principle of ends, then I can without contradiction adhere to the principle of ends and the doctrine of innocent until proven guilty, even though I admit that from the perspective of theoretical reason it is an open question as to whether free will exists.

Shouldn't the dictates of theoretical reason be given greater weight in our thought and action than the dictates of practical reason?

Before concluding let's consider one last objection to my argument. A critic might note how I am suggesting that practical reason might lead us to the acceptance of a moral principle—the principle of ends—which, as I argue, presumes the existence of free will. The critic might also note that I concede that in the domain of theoretical reason there is no epistemic justification for this belief in free will; rather, it is an open question whether it exists. Here the critic might assert that theoretical reason—what reason dictates regarding science, math, logic, metaphysics, etc.—should have primacy of place in our thinking and how we live our lives. Thus, if a moral principle entails adoption of a belief that theoretical reason cannot support, such as a belief in free will, then we should not embrace the moral principle.

In response, I want to first note that when a moral principle conflicts with something that we clearly know to be true in the realm of theoretical reason then we should reject the moral principle. But such is not the case regarding the issue before us. Rather, my view, like that of most libertarians, is that it is an open question whether free will exists; there's not a whole lot of evidence for its existence or for its nonexistence. As I've argued, rational acceptance of the principle of ends involves a presumption of the existence of free will. If I have good moral reasons to adopt the principle of ends, then I don't see why I should withhold from adopting it just because theoretical reason provides no sufficient evidence of free will. To suggest that I should wait until there is theoretical proof of free will is to give theoretical reason an exalted status without good reason for doing so.

Conclusion

In this essay, I hope to have shown how it is that libertarians can be morally justified in holding persons responsible for their actions while admitting they lack epistemic justification for their beliefs in libertarian free will. In concluding, I would note that the libertarian perspective might sometimes play a role in leading some people to be overly harsh in their blaming and punishing of persons. When we view persons as possessing a libertarian free will that gives them ultimate responsibility for their character and actions,

it can easily lead us to think wrongdoing merits equal levels of blame and punishment directed towards anyone who has committed the same offense. But here we have to be careful. Just because two persons have libertarian free will, it does not mean their life circumstances and the pressures and temptations they face are the same. It is unjust not to take these matters into consideration when levying blame and punishment upon persons. It may well be that one thief or drug dealer deserves less punishment than another even if they've committed the same crimes and both have libertarian free will. This is because we should in blaming and punishing acknowledge that one might have faced greater pressures and temptations, making it more difficult for him to act rightly. These considerations are perfectly consistent with a libertarian perspective and they must be kept in mind by libertarians so as to avoid actually being or becoming hardhearted.

References

Balaguer, Mark. 1999. "Libertarianism as a Scientifically Reputable View." *Philosophical Studies* 93 (2): 189–211.

Balaguer, Mark. 2010. *Free Will as an Open Scientific Problem.* Cambridge: MIT Press.

Chisholm, Roderick. 1976. *Person and Object*. LaSalle, IL: Open Court.

Clarke, Randolph. 1995. "Toward a Credible Agent-Causal Account of Free Will." In *Agents, Causes, and Events*, edited by Timothy O'Connor, 201–215. New York: Oxford University Press.

Clarke, Randolph. 2003. *Libertarian Accounts of Free Will.* New York: Oxford University Press.

Double, Richard. 2002. "The Moral Hardness of Libertarianism." *Philo* 5 (2): 226–234.

Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas. 2007. *Four Views on Free Will*. Malden, MA: Blackwell.

James, William. (1884) 1956. "The Dilemma of Determinism." In *The Will To Believe: Human Immortality*, 145–183, 1956. New York: Dover.

Kane, Robert. 1996. *The Significance of Free Will.* New York: Oxford University Press.

Kane, Robert. 2007. "Libertarianism." In *Four Views on Free Will*, edited by John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, 5–43. Malden, MA: Blackwell.

Kane, Robert. 2011. "Rethinking Free Will: New Perspectives on an Ancient Problem." In *The Oxford Handbook of Free Will*, edited by Robert Kane, 381–404. New York: Oxford University Press.

Kant, Immanuel. (1785) 1993. *The Groundwork of the Metaphysics of Morals.* Translated by James W. Ellington. Indianapolis: Hackett.

Kant, Immanuel. (1788) 1927. *The Critique of Practical Reason*. Translated by T.K. Abbott. New York: Longman, Green.

O'Connor, Timothy. 1995. *Agents, Causes, and Events.* New York: Oxford University Press.

O'Connor, Timothy. 1995a. "Agent Causation." In *Agents, Causes, and Events*, edited by Timothy O'Connor, 173–200. New York: Oxford University Press.

O'Connor, Timothy. 2000. *Persons and Causes: the Metaphysics of Free Will*. New York: Oxford University Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Rowe, William. 1995. "Two Concepts of Freedom." In *Agents, Causes, and Events*, edited by Timothy O'Connor, 151–171. New York: Oxford University Press.

Taylor, Richard. 1966. *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

Waller, Bruce. 2011. *Against Moral Responsibility.* Cambridge: MIT Press.

Journal of Cognition and Neuroethics

Libet, Free Will, and Conscious Awareness

Janet Levin
University of Southern California

Biography
Janet Levin is Associate Professor of Philosophy at USC. Her research interests include philosophy of mind, philosophy of psychology, and epistemology.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Levin, Janet. 2015. "Libet, Free Will, and Conscious Awareness." *Journal of Cognition and Neuroethics* 3 (1): 265–280.

Libet, Free Will, and Conscious Awareness

Janet Levin

Abstract

In a series of well-known experiments, Benjamin Libet fits subjects with electrodes that monitor their brain activity, and instructs them to decide whether or not to flex their wrists at various times during a certain interval and then follow through. And—notoriously—he finds that the subjects' wrist flexings are preceded by the occurrence of a 'readiness potential' (RP) that begins about 400ms *before they report any conscious inclination or wish to act*. Therefore, Libet argues, these wrist flexings do not arise from the subjects' free will. There have been numerous attempts to dispute Libet, and argue that his subjects' conscious wishes or inclinations can be regarded as the causes of their actions—and I find many of these arguments compelling. Here, however, I question the connection between conscious motivation and freedom of action, and argue that behavior produced by wishes or inclinations of which we are not consciously aware can often be viewed as sufficiently up to us, or under our control, to count as free action. On the other hand, as I argue, we may need to be consciously aware of our motivations to be held *morally responsible* for what we do. And this, I suggest, has some potentially interesting implications for our common views about the relation between free will and moral responsibility.

Keywords

Libet, free will, conscious experience, moral responsibility

In his well-known experiments that purport to show that we have less free will than we may think, Benjamin Libet (1985, 2011) fits subjects with electrodes that monitor their brain activity, and instructs them to decide whether or not to flex their wrists at various times during a certain interval and then follow through. And—notoriously—he finds that the subjects' wrist flexings are preceded by the occurrence of a 'readiness potential' (RP) that begins about 400ms *before they report any conscious inclination or wish to act*. Therefore, Libet argues, these wrist flexings do not arise from the subjects' free will. It may *seem* to the subjects that they are consciously willing to flex their wrists, but this is merely an illusion.

On the other hand, Libet argues, the data show that if the subjects become aware of their inclinations to flex and consciously 'veto' such inclinations, their subsequent actions *are* directly initiated by a conscious process, and thereby *do* arise from their free will—at least insofar as their conscious vetoes are not themselves determined.[1] In short, on Libet's

1. As V. S. Ramachandran (and subsequently many others) have put it, for Libet there is no freedom of will, but only freedom of 'won't.' See, however, Lau et al (2007) for skepticism about whether vetoing is a

view, being initiated by a conscious wish or willing is *necessary* for an action to originate from a subject's free will—though perhaps not *sufficient*.

There have been numerous attempts to show that, contrary to Libet's suggestions, a subject's conscious wish or inclination to flex *can* be regarded as the cause of the flexing. For example, some (e.g., Roskies, 2011, Dennett, 2003, and Mele, 2011) argue that the activation of the RP is merely the *lead-up* to a subject's conscious willing, and not the willing itself; others (e.g., Roskies, 2011) suggest that the activation of the RP may well *be* the subject's conscious willing, which precedes (by a few ms) the subject's *report* (or even *conscious awareness*) that it has occurred. Alternatively, Horgan (2011) argues that, even if the initiation of the RP truly precedes the conscious willing to act (and not just the conscious awareness of that wish or inclination), that conscious state can be regarded as the *sustaining cause* of the implementation of a standing intention to act, and thus does not threaten the veridicality of the experience of conscious will.[2] I find many of these arguments compelling. However, they all focus on challenging the claim that the actions in question do not originate from the agent's *conscious wish or inclination to act*—and this implies that being initiated by a conscious willing is *necessary* for an action to be free.

Here, however, I want to question the connection between conscious motivation and freedom of action, and consider whether behavior produced by wishes or inclinations of which we are *not* consciously aware can nonetheless be viewed as sufficiently up to us, or under our control, to count as free action. I will argue that the answer, at least sometimes, may be 'yes,' and thus that the focus on whether Libet's findings undermine the view that our actions are produced by conscious motivation may be less relevant to determining whether we have free will than is often assumed.

On the other hand, I also will consider the relation between conscious motivation and *moral responsibility*; in particular, whether (we think) an agent can be blamed or praised for doing something if she has no conscious awareness of a decision, or wish, or inclination to do so. And I will argue that in this case the answer, more often, may be 'no.' The upshot of these considerations will be that there is reason to think that although an agent's conscious decision (or wish or inclination) to do A may not be necessary for

conscious activity, and Mele (2013c) for discussion of these, and other, findings.

2. In a different attempt to counter Libet, still others (e.g., Roskies, 2011, Mele 2013a,b, 2014) argue that even if Libet is correct to claim that his subjects' wrist-flexings do not arise from free will, this conclusion may not generalize to actions that are the products of more extensive deliberations in which conscious decision plays an important *contributing* role.

A to be *free*, (we tend to think) it may be necessary for the agent to be held *morally responsible* for doing A. And this has some potentially interesting implications for our common views about the relation between freedom of action and moral responsibility.

To think about these issues, let's consider a more 'real life' scenario that, at least at first glance, has affinities with Libet's wrist-flexing experiments:

Suppose I'm a professional diver, and I practice for at least three hours every day. I go up to the top of the high dive—without thinking much about what I'm doing—position myself, and dive directly into the water. I swim to the edge of the pool, then climb up and do it again—and again. Occasionally, however, something doesn't seem quite right: some debris in the water, a kid who's swimming toward the diving board, or something I can't quite put my finger on—and I don't make the dive, or I dive in a different direction. When everything is going well, it doesn't seem like I'm *consciously willing* to dive (or to dive in the particular direction that I do) or even that I'm consciously aware of an inclination to do so. Maybe, given that I told my coach that I'd indicate when I would attempt a dive, I give a thumbs up as I'm about to leave the platform. But sometimes I forget and just do it.

This vignette, of course, has at least some commonalities with the situation of Libet's subjects: when things go well, my indication of intention (thumbs up) seems equally after the fact. And I suspect that if there had been (waterproof) electrodes affixed to my head, my brain activity would look similar to that of Libet's subjects.[3]

As we know, Libet contends that in the situations in which his subjects decide to flex their wrists and do so, they don't act freely; only acts initiated by a conscious veto can be the products of free will. But is it so obvious that when I make a straightforward dive in the situation described *I* don't act of my own free will? I suspect that this may seem less clear. And this is so, it seems, even though my diving—as in many other cases of so-called *skill exercise*—may seem *more* 'automatic,' *less* governed by anything like conscious will, than the wrist flexing of Libet's subjects. Indeed, it's not clear that my diving in a different direction (or not diving at all) when things seem sketchy is best regarded as the result of a 'conscious veto' of a wish or inclination; these actions seem pretty automatic as well, at least if I'm truly a skilled diver whose training has made

3. After all, when I do take the dive, there must be *something* going on in my brain that precedes my action prior to my giving a thumbs-up to my coach, just as there is something going on in the brains of Libet's subjects before they express their inclination to flex their wrists.

for flexibility of response. But here too—or so it seems—my action (or refraining from action) seems to be something that I did freely, something that was up to me.[4]

Libet, no doubt, would disagree. But is it clear that this is the right verdict? In what follows, I want to address three questions. First, are there conditions under which seemingly automatic actions like diving—actions not obviously caused by a conscious decision—can nonetheless be products of free will; second, do these conditions require that the agent have *any* sort of conscious awareness—and if so, awareness of *what*; and third, do we have the same views about the relation between conscious awareness and moral responsibility?

First, let's examine the relation between freedom and automaticity. Some theorists argue that automaticity—even the sort exemplified in skill exercise like piano playing or diving—is incompatible not only with *free* agency, but with any sort of intentional agency at all.[5] But this view is far from universal. Among the dissenters is Mele (2011), who argues that intentional action can occur without a conscious decision to act in that way when the action is routine, and suggests (26) that if subjects have a *conditional intention* to do something when they feel like it, and act forthwith, then their actions can count as intentional. Wayne Wu (2013a) goes further, and argues (258) that, in 'normal action' produced by intention, 'intentions are persisting nonphenomenal states of subjects that coordinate and constrain one's meandering through behavioral space.' I find these views about agency plausible, and if they are right, then it seems that my diving and Libet's subjects' flexings can at least sometimes be intentional actions even if not produced by an explicit conscious decision. But I want to go even further and question whether an act can be intentional not just if the agent has no conscious intention, conditional or not, to perform that act, but whether conscious awareness of *anything* at all is required for agency.

4. There are, of course, many examples of (what seems to be) free action without conscious choosing that aren't examples of skill exercise, such as pulling out one's wallet to pay the check, or getting up to answer the ringing doorbell. See Vihvelin (2013, 6.3) for other good examples of choice without conscious choosing, and discussion.

5. See, for example, G. Strawson's (2003) argument that automaticity undermines agency. Also, consider Nadelhoffer (2011, 178), 'The agential threats that we will be examining here are ultimately fueled by the fact that the conscious mind exercises less control over our behavior than we have traditionally assumed. It is this deflationary view of conscious volition that is potentially agency undermining.' He continues (183), 'So whereas Libet's view [involving the possibility of 'veto power'] merely shrinks the domains over which we exercise control, Wegner seemingly leaves the conscious mind out of the causal loop altogether…As such, Wegner's view…represents a global and not merely partial agential threat.'

Now one may think that this suggestion is crazy: for an individual to perform a bona-fide intentional action, one may think, she must at least have conscious awareness of *something*, perhaps the environmental conditions in which she does what she does. After all, how could a piano player or a diver possibly play the right notes or dive in the right direction if she were not consciously aware of such things as the location of the piano keys or the water?

But is this connection so clear? Consider, for example, David Chalmers's (1995) distinction between the 'hard' and the 'easy' problems of consciousness. The hard problem, according to Chalmers, is to give a satisfying explanation of why it is that 'when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C…[that is] why there is something it is like to entertain a mental image, or to experience an emotion.' This problem, he contends, is different from the 'easy' problems of providing a satisfying explanation of phenomena such as:

(1) the ability to discriminate, categorize, and react to environmental stimuli
(2) the integration of information by a cognitive system
(3) the reportability of mental states
(4) the ability of a system to access its own internal states
(5) the deliberate control of behavior
(6) the difference between wakefulness and sleep.

These problems, Chalmers acknowledges, are hard, but hard in a different way, in that even though it may take time and effort to solve them, there is 'no real issue about whether these phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms.'[6] And the reason is that it seems that all we have to do to provide a satisfying explanation of those phenomena is to (i) get clear about the way these processes function—and then (ii) look around (in the brain and body) for some mechanism that performs or implements this function. Once we find such mechanisms, the question seems closed; we're done.

6. He continues, 'To explain access and reportability, for example, we need only specify the mechanism by which information about internal states is retrieved and made available for verbal report. To explain the integration of information, we need only exhibit mechanisms by which information is brought together and exploited by later processes. For an account of sleep and wakefulness, an appropriate neurophysiological account of the processes responsible for organisms' contrasting behavior in those states will suffice. In each case, an appropriate cognitive or neurophysiological model can clearly do the explanatory work.'

But this is not so, Chalmers contends, for the how and why of conscious experience. Even if we were to have a comprehensive scientific explanation of (1)-(6), we would have no satisfying explanation of why a creature with these capacities has the conscious experiences it has—or even why it has any conscious experiences at all. Indeed, he argues, it seems possible that there could be creatures just like ourselves with respect to (1)-(6) but with no phenomenally conscious experiences.[7]

Now if the idea of such a creature is coherent, then it's conceivable that an individual could perform a wide variety of behaviors produced by sophisticated cognitive processes even though it doesn't have conscious awareness of environmental conditions—let alone any *experience of agency* that some take to a necessary antecedent of an intentional action. And this suggests that it's at least conceivable that a creature could be an intentional agent without being conscious. Indeed, we don't have to invoke such Chalmerisan constructs to make the point, since there is at least some empirical evidence (from Kentridge et al, 1999) that the well-known blindsight patient GY can discriminate objects in his blind hemi-field. If this is so, then conscious awareness of one's environment may be merely a *de facto*, but not *in principle*, requirement for intentional agency.[8]

7. For example, he continues, 'we can imagine that right now I am gazing out the window, experiencing some nice green sensations from seeing the trees outside, having pleasant taste sensations through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder' and has a counterpart wrt (1)-(6) that 'will be perceiving the trees outside in the functional sense, and tasting the chocolate in the psychological sense, [and will be] awake, able to report the contents of his internal states, able to focus attention on various places…[but] none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel' (1996, 94-5). Now, I have characterized Chalmers as arguing that it is imaginable that there are creatures that possesses capacities (1)-(6), but have no conscious experiences. But he makes a considerably stronger claim, namely that we can conceive or imagine creatures that are *our exact molecular duplicates* but have no conscious experiences. This claim is more controversial than the one I am considering, but for the purposes of this discussion the weaker claim is strong enough. (In their discussion of 'the zombie challenge' in the introduction to their (2013) collection, Vierkant, Kiverstein, and Clark are clear about this distinction.)

8. See Wu (2013a), who discusses these findings in service of his view that attention can be 'selection for action' even if not conscious: 'we would fail to fully capture an essential psychological capacity were we to restrict talk of attention to just these conscious forms. For the very capacity for action requires that the agent exhibit a striking form of attunement to the world so as to guide her behavior, and much of this attunement is in a way subterranean to consciousness even if it is not subpersonal. Responsiveness to the world, in action, precisely involves a way of attending to the world, more often unconscious than not.' See also Wu (2013b).

Nonetheless, an action can be intentional without being free, and so we need to consider whether *free* intentional action may require some sort of conscious awareness.[9] In the literature on free will and free agency, theorists have suggested a variety of conditions that must be met for an action to arise from true freedom of the will. In much of this literature the guiding question is whether these conditions could be met in a deterministic universe. But I want to ignore that question and focus solely on whether one can meet any of these conditions without conscious awareness.

Consider, first, Harry Frankfurt's well-known (1971) account of what it takes for an act to proceed from a 'will that is free,' namely, that (i) I act freely; that is, *I do what I want*, and (ii) I act of my own free will; that is, I *want* to act (or I approve of my acting) on that desire, and finally, (iii) *if I hadn't wanted to act on that desire, it wouldn't have been the one that caused my action*.

This characterization, it seems, reflects some intuitive distinctions we may be inclined to make in variations of the diving case. Suppose I'm preparing to dive, as usual, and a little kid swims under the board—and I make my usual dive and hit him. Did I do *this* of my own free will? Well, it seems that we would have to consider a number of things: Did I notice the kid before I started the dive? If I did notice the kid, was my dive too far along for me to stop or change my direction if I had wanted to? Presumably, the answer to the first question will sometimes be 'no,' and the answer to the second will sometimes be 'yes.' In these cases it does indeed seem that my action is not the product of my free choice. On the other hand, things could be different. Suppose that I did notice what was going on, but made the dive anyway because I wanted to. Suppose I also approved of that desire (because this was my only chance for a gold medal!), and also that I would have stopped or changed direction if I had wanted to. In that case, it seems—at least arguably—that I *did* act of my own free will. But what sort of *consciousness*, if any, is required for me to have acted in this way?

Consider Frankfurt's Condition (i): I did what I wanted to do; that is, my dive was caused by my wanting to dive (or it was produced, at least, by a 'conditional intention' to dive under certain circumstances). Now if, as Chalmers contends, an unconscious creature is capable of the 'deliberate control of behavior,' then it seems that it could

9. Even Wu (2013a), who is sanguine about the possibility of agency without conscious awareness, acknowledges that *free* agency may require something more. See his (2013, 258): 'There is no denying we are often moved to act, and on the antecedents of action rest important questions about the rationality, morality, and freedom of our actions. But all thee are higher-order properties of agency. Agency itself is an internal feature of certain processes….'

do what it wants in the same way that I did. Frankfurt's second condition, however, may seem harder to meet, since it involves second-order desires. That is, to act 'of one's own free will,' rather than merely 'acting freely,' requires that the agent want to act on her effective desires (or at least approve of so acting), and this requires a capacity for *reflecting on* the (1st order) desires she has. After all, for Frankfurt, non-human animals may be capable of free action in the sense that nothing prevents them from doing what they are inclined to do, but they are incapable of acting on their own free will because they do not possess the reflective capacities that would allow then to take a stand on the desires (or inclinations) they have. And surely, one might think, to meet this 'hierarchical' condition on freedom one must be *consciously aware* of one's desires.

But is the connection between having reflective capacities of this sort and being (phenomenally) conscious so clear? Consider once again Chalmers's imagined creatures: creatures that meet conditions (1)-(6), but with no (phenomenally) conscious experience. As mentioned above, such creatures are supposed to have the capacity not only to behave the way we do, but also (among other things) to 'discriminate, categorize, and react to environmental stimuli' and to access and report on their own mental states. It seems, therefore, that such creatures could have the reflective capacities required to meet all Frankfurt's conditions for having a will that is free. The same, it seems, can be said about views (e.g. Watson, 1975) that take free actions to be those compatible not with one's second-order desires, but with one's *values*, as long as values can be characterized in some naturalistic way.[10]

The same questions, moreover, can be raised for other contemporary characterizations of actions that arise from an agent's free will. Kadri Vihvelin, in her recent (2013) book, contends that an action arises from an agent's free will only if the agent could have done otherwise, where this requires having both the *ability* and *opportunity* to have done so. Vihvelin gives a subtle and rich characterization (176) of what counts as the sort of ability (or abilities) relevant to our concerns, namely, that one have 'an intrinsic disposition to do X in response to the stimulus of *one's trying to do X*.' Moreover, she continues, '[a]ll that's required for trying is that you acquire—somehow or other—an effective desire or intention; that is, a desire or intention that is causally effective [in certain specified ways].'[11] Here too, however, it's not clear that possessing (to coin a

10. Indeed, even outside the realm of Chalmersian beings, it is easier to think of situations in which people are blind to what they *really* consider to be important; self-deception abounds in the domain of value.

11. Once again, both Frankfurt and Vihvelin go on to argue that acts can meet these conditions in a deterministic universe, but I'm not interested in evaluating that contention here.

phrase) Vihvebilities requires some sort of (phenomenally) *conscious awareness*—either of one's own motivations or the environmental conditions that afford opportunities to act. That is, it's not clear why an unconscious creature that meets conditions (1)-(6) could not have these abilities too.

Yet another characterization of actions arising from free will is that the actions are *reasons-responsive*. Does this require phenomenal consciousness either of one's own mental states, or items in one's environment? Let's look more closely at what it takes for an action to be reasons-responsive. One possibility is that the agent 'acts for reasons' in the sense that her action must be caused not just by a desire, but by a *rational* desire—that is, a desire that (either) coheres sufficiently with the rest of her desires (and beliefs and perhaps values) or with the desires, beliefs, and values that meet certain independent conditions of rationality. Either way, it's not clear that an unconscious creature that meets conditions (1)-(6) could not act for reasons in this sense, even though it is completely (phenomenally) unconscious.

Another way to think about reasons-responsiveness is that an agent must be able to modify her actions if she *recognizes that there are reasons to do so*—whether these are beliefs and desires that she has suppressed or otherwise not yet noticed, or environmental conditions that turn out to be different than expected. But once again it's unclear why an unconscious creature that meets conditions (1)-(6) couldn't do *that*.

The upshot, thus, is that it seems that a creature without (phenomenally) conscious awareness could be capable of acting freely, or acting from its own free will, on any of these causal, structural, or counterfactual accounts of freedom.[12]

However, even if it's possible for such creatures to exercise free will, *we* are not beings of this sort; we have conscious access both to our own experiences and the world around us, and it seems that many of our actions, including those studied by Libet, arise from a conscious wish or inclination. *We* have the experience of agency (Horgan, 2011) or of conscious will (Wegner, 2002). And if these conscious inclinations do not in fact initiate our actions, but our actions nonetheless count as arising from free will, then

12. Of course there are other views of what's required for free will; for example, that the action be the product of *agent causation*; that is, that it be caused not by events such as beliefs, desires, or intentions, but by the agent herself. It's a matter of debate whether the idea of agent causation is coherent; but it's not clear why a creature of the sort we've been discussing couldn't be the cause of its actions in just the way that a reasonable theory of agent causation demands. Granted, often people object to the possibility that a robot or similar machine could have free will—but this is usually on the grounds that such a creature would be acting according to the way it was programmed, and not because there are questions about whether such creatures could be phenomenally conscious.

our experiences of agency, or of the efficacy of our conscious will, must be regarded as illusory—just as Libet, Wegner, and many others have argued, and this would be disturbing in itself.

Perhaps this is so (though some, e.g. Paglieri, 2013, and Gallagher, 2013, have argued that the phenomenology of agency is not as robust as Libet and Wegner suggest).[13] In any case, the illusoriness of the experience of agency, if it obtains (remember that there is dispute about whether Libet's data show that his subjects' actions do not arise from their conscious inclinations), is compatible with the *existence* of human free will. And if our experiences of agency are systematically erroneous, then they join a fairly large club. It's not unusual for us to be wrong about the causes of a variety our actions—and about other phenomena presented in introspection.

To be sure, it would be disturbing to think that our actions may be caused by some sort of nefarious manipulator while it seems to us that they are caused by our conscious decisions. However, what makes this possibility so disturbing, I suggest, is not that that the causes of our actions are not conscious, but that they are motivations that we would not want to be effective, or that diverge from our values, or are inconsistent with the motivations we identify with, or take to reflect our real or 'deep' selves. But, according to the best-known views about what makes an action free, these would not be cases in which we err about the motivations for our free actions, but rather cases in which our actions are not in fact free.

However, I suspect that things may seem different for attributions of *moral responsibility*. It's harder, it seems, to blame or praise people for what they do if they have no (phenomenally) conscious experience either of their desires and inclinations (be they conflicting or consistent) or of the environment that is prompting them, given those desires and inclinations, to act in certain ways.

Consider the sorts of things we say to excuse people from blame (or question whether they are being legitimately praised): 'She wasn't aware of what she was doing'; 'He didn't recognize the difference between right and wrong.' These excuses, it seems, carry with them *not only* the suggestion that the agent wasn't responsive, in some way, to environmental (and internal) exigencies, but also that the agent did not have

13. As Paglieri puts it (2013, 136), 'There is no proof that free actions are phenomenologically marked by some specific 'freedom attribute…and thus this non-existent entity cannot be invoked to justify our judgments on free will and agency, be they correct or mistaken.' He goes on to suggest that we regard our acts as free by default, and only consider that they are not free if we have countervailing evidence, such as the experience of coercion. Gallagher (2013) argues that the sense of agency can sometimes arise after retrospective reflection, and often has a social dimension.

a robust conscious awareness of them, the sort that comes with the ability to adopt another person's subjective situation, or point of view. To be able to do this, of course, requires that one *have* a point of view, in the sense of having phenomenally conscious experiences of (and, perhaps more important, emotional responses to) the things that are the causes and effects of one's actions. If so, then moral agency may require (phenomenal) consciousness. That is, it may seem that unconscious creatures (with the relevant functional capacities) may be able to be intentional agents, indeed free agents—but not moral agents.[14]

This view has at least some support in the literature on moral responsibility. For example, T.M. Scanlon (1998, 281) argues that although a computer could be regarded as responsible 'in a causal sense for the processes it governs'…we would not 'regard it as "responsible" in the sense responsible for moral blame.' And this, he continues (282) is 'because computers, even very sophisticated ones, …are not conscious.'[15]

If acts produced by unconscious motivations could nonetheless be free, however, this would not be the only case in which actions could be regarded as free but not blameworthy (or praiseworthy). Think, for example, of actions that are coerced: Someone holds a gun to your head and says 'Your money or your life'—and you hand over your money. Although some claim that you don't do so freely, it is equally intuitive to hold

14. In addition, there are views that contend that having certain *emotional* responses is necessary for recognizing others as bona-fide moral agents who can legitimately be blamed and praised for what they do. See P.F. Strawson (1962).

15. Indeed, Scanlon goes further and argues (282) that 'it is crucial to a creature's being a rational creature that conscious judgment is one factor affecting its behavior. Computers, even very sophisticated ones, do not strike us as moral agents or rational creatures, partly because we believe that they are not conscious at all—and that there is no such thing as how reasons, or any other things, seem to them.' (I am grateful to Pamela Hieronymi for calling these passages to my attention.) Scanlon's mention of how things *seem* to an agent suggests that he means by 'consciousness' just what Chalmers takes to be independent of the capacities (1)-(6) discussed in the text. On the other hand, Michael S. Moore (2011, 223, col. 1) argues that even in the absence of 'phenomenal' (that is, Chalmersian) consciousness, the possession of 'dispositional' consciousness of one's motivation—that is, 'the ability to direct attention and to state that of which one is conscious, abilities that seem included in Chalmers's (1)-(6)—may be sufficient for moral responsibility. Nonetheless, Moore (223, col. 2) argues (with respect to Freudian explanation) that 'Although there may be truly unconscious agency that is nonetheless the agency of a person, that person's responsibility is not increased by virtue of such truly unconscious actions, intentions, or tryings.' It's not clear, however, just what counts as 'truly unconscious agency,' and whether it could be possessed by a creature that satisfies any of the Frankfurtian or other well-known conditions for the possession of free will. In any case, Moore goes on to argue that, in the Libet cases, agents can be regarded as acting because of willing that is conscious in both the phenomenal and dispositional sense.

that you *do* make a free choice and thereby act freely, but since the costs and benefits of your available alternatives have been manipulated in certain ways, you shouldn't be held responsible for what you do (or choose). Your actions could be 'yours' or 'up to you' metaphysically, but not in ways that make a difference to praise, blame, or other sorts of moral evaluation.

If moral responsibility does not follow automatically from freedom in cases of coercion, then perhaps it does not follow from freedom in Chalmersian (or blindsight) cases, either. Indeed, the separation of questions about free will from questions about moral responsibility may make certain discussions of Libet's results more intelligible. For example, in his (2013b), Mele considers a situation in which an agent consciously deliberates among alternatives but makes no decision about them, and takes up the deliberation days later, still feeling 'unsettled' about what to do. Mele suggests that if the agent were to find out that she *unconsciously decided to do A after her initial period of deliberation*, then she may doubt that she did A freely. He argues, however, that this case is significantly different from a case in which the time lag between the action and the prior decision (made after conscious deliberation) is much shorter, on the order of half a second. In this case, Mele argues, the time lag is not threatening, since we can think that our detection of the decision merely 'lags a bit behind the actual decisions.' He acknowledges that it may be odd to hang freedom of action on the time interval between decision and action in this way, and suggests (786) that the presence of a long time lag between decision and awareness prompts skepticism about whether conscious deliberation did in fact play a significant role in the production of the decision. It may make more sense, however, to argue that the time lag makes a difference not to the determination of whether the action was free, but to the attribution of *moral responsibility* to the agent.[16]

Now I myself do not have firm views about whether moral responsibility requires conscious awareness of one's motives (or anything else); I'm willing to consider the possibility that unconscious agents (who meet the conditions for freedom) are morally responsible as well. But perhaps the recognition that there may be a gap between freedom and responsibility will make it easier to get a clear-eyed account of what is required for freedom—and, perhaps more important, a clear-eyed account of the aims of praise and punishment. If we separate considerations of whether an individual is conscious from questions about whether she is free, then we may be able to get a better

16. Moreover, it can help to make sense of Wegner's claim, highlighted by Dennett (2003, 242) that '[i]llusory or not, conscious will is the person's guide to his or her own moral responsibility for action.'

hold not just on questions about freedom, but also on questions of moral responsibility, and therefore, this may be a view that is worthy of further discussion.[17]

17. The idea that we should firmly separate considerations of the aims of praise and punishment from consideration of whether an agent acts freely is familiar from the work of hard determinists, such as Holbach, who argues that this is required if people *cannot* exercise free will. It may be required as well for those who hold that humans can act freely in a wider variety of circumstances than we had initially thought. In addition, the separation of considerations about whether an individual is conscious from considerations about whether she is free—or could even be an agent—may have salutary effects for further theorizing about other phenomena that seem to be tied to action or intentional agency, such as attention.

References

Block, Ned. 1995. "On a Confusion about the Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–87.

Chalmers, David. 1995. "Facing up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200-219.

Clark, Andy, Julian Kiverstein, and Tillmann Vierkant. 2013. *Decomposing the Will*. New York: Oxford University Press.

Dennett, Daniel C. 2003. *Freedom Evolves*. New York: Penguin Group (Viking).

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20.

Gallagher, Shaun. 2013. "Ambiguity in the Sense of Agency." In *Decomposing the Will*, edited by Andy Clark, Julian Kiverstein, and Tillmann Vierkant, 118–135. New York: Oxford University Press.

Horgan, Terry. 2011. "The Phenomenology of Agency and the Libet Results." In Conscious Will and Responsibility: A Tribute to Benjamin Libet, edited by Walter Sinnott-Armstrong and Lynn Nadel, 159–172. New York: Oxford University Press.

Libet, Benjamin. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action." *Behavioral and Brain Sciences* 8 (4): 529–66.

Libet, Benjamin. 2011. "Do We Have Free Will?" In *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 1–10. New York: Oxford University Press.

Mele, Alfred. 2011. "Libet on Free Will: Readiness Potentials, Decisions, and Awareness." In *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 23–33. New York: Oxford University Press.

Mele, Alfred. 2013a. "Free Will and Neuroscience." *Philosophical Exchange* 43 (1): Article 3.

Mele, Alfred. 2013b. "Unconscious Decisions and Free Will." *Philosophical Psychology* 26 (6): 777–789.

Mele, Alfred. 2013c. "Vetoing and Consciousness." In *Decomposing the Will*, edited by Andy Clark, Julian Kiverstein, and Tillmann Vierkant, 73–86. New York: Oxford University Press.

Mele, Alfred. 2014. *Free: Why Science Hasn't Disproved Free Will*. New York: Oxford University Press.

Nadelhoffer, Thomas. 2011. "The Threat of Shrinking Agency and Free Will Disillusionism." In *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 173–188. New York: Oxford University Press.

Paglieri, Fabio. 2013. "There's Nothing Like Being Free: Default Dispositions, Judgments of Freedom, and the Phenomenology of Coercion." In *Decomposing the Will*, edited by Andy Clark, Julian Kiverstein, and Tillmann Vierkant, 136–159. New York: Oxford University Press.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Sinnott-Armstrong, Walter, and Lynn Nadel. 2011. *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. New York: Oxford University Press.

Strawson, Galen. 2003. "Mental ballistics or the involuntariness of spontaneity." *Proceedings of the Aristotelian Society* 103: 227–256.

Strawson, P.F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.

Vihvelin, Kadri. 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wu, Wayne. 2011. "Attention as Selection for Action." In *Attention: Philosophical and Psychological Essays*, edited by Christopher Mole, Declan Smithies, and and Wayne Wu, 97–116. New York: Oxford University Press.

Wu, Wayne. 2013a. "Mental Action and the Threat of Automaticity." In *Decomposing the Will*, edited by Andy Clark, Julian Kiverstein, and Tillmann Vierkant, 244–261. New York: Oxford University Press.

Wu, Wayne. 2013b. "The Case for Zombie Agency." *Mind* 122 (485): 217–230.

Journal of Cognition and Neuroethics

Experimental Philosophy, Robert Kane, and the Concept of Free Will

J. Neil Otte
University at Buffalo, The State University of New York

Biography

J. Neil Otte is presently in the Ph.D. program at the University at Buffalo (SUNY) and works in moral psychology, cognitive science, and the history of philosophy. For six years, he was an adjunct lecturer in philosophy at John Jay College of Criminal Justice in New York City. In the last three years, he has been an organizer of the Buffalo Annual Experimental Philosophy Conference, an organizer for a conference on sentiment and reason in early modern philosophy, and has worked as an ontologist for the research foundation, CUBRC.

Acknowledgements

The author wishes to thank Gunnar Björnsson, Gregg Caruso, and Oisín Deery for their supportive comments and conversation during the Free Will Conference at the Insight Institute of Neurosurgery and Neuroscience in Fall of 2014, as well as conference organizer Jami Anderson.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Otte, J. Neil. 2015. "Experimental Philosophy, Robert Kane, and the Concept of Free Will." *Journal of Cognition and Neuroethics* 3 (1): 281–296.

Experimental Philosophy, Robert Kane, and the Concept of Free Will

J. Neil Otte

Abstract

Trends in experimental philosophy have provided new and compelling results that are cause for re-evaluations in contemporary discussions of free will. In this paper, I argue for one such re-evaluation by criticizing Robert Kane's well-known views on free will. I argue that Kane's claims about pre-theoretical intuitions are not supported by empirical findings on two accounts. First, it is unclear that either incompatibilism or compatibalism is more intuitive to nonphilosophers, as different ways of asking about free will and responsibility reveal different answers. Secondly, I discuss how a study by Josh May supporting a cluster concept of free will may provide ethicists with a reason to give up a definitional model, and I discuss a direction future work might take. Both of these objections come from a larger project concerned with understanding the cognitive mechanisms that people employ when they make judgments about agency and responsibility—a project that promises not only to challenge contemporary philosophy, but to inform it.

Keywords

Free will, experimental philosophy, compatibalism, incompatibilism, moral cognition, Robert Kane

Introduction

Studies that employ experimental method to examine non-philosophers' concept of free will have been going on for some time now, and trends in the literature are now forming, giving us reason to re-evaluate some contemporary positions. As an example of how this work lends support to a re-evaluation of the contemporary literature, I take as my focus the influential work of Robert Kane. In this essay, I argue that evidence that nonphilosophers are compatibilists or incompatibilists is presently not forthcoming, that evidence showing that moral judgment affects judgments about happiness, the mental states of others, and knowledge (Beebe & Buckwalter 2010; Knobe ; Phillips, Nyholm, & Liao) might give us pause about accepting Kane's account of self-forming actions, and that work by Josh May supports rejecting a classical account of the concept of free will in favor of a prototype or exemplar account.

I. Kane and the Concept of Free Will

Philosophers often appeal to the ordinary conception of free will. The reason for this is that we're not just interested in any old conception of free will from which we could avoid hard problems and deduce and defend easy platitudes; rather, we're interested in the concept that people actually use, the concept that people employ when they judge an action to have been freely performed, and frequently, the judgment that entails that an action is capable of being praised or blamed. Philosophers who find this train of thought compelling argue that when we judge whether incompatibilism or compatibalism is true or false, it must be judged according to the conception of free will as it is found in the wild.

I use 'concept' here as it is used most often in psychological literature, and not principally as it is often used in other fields like computer science and philosophy. Thus, I take the concept of free will to be the body of knowledge about free will that is characteristically used in the cognitive processes that underwrite our judgments of free will.[1] Popular models of such concepts include definitions, theory-theories, exemplars, and prototypes; and experimental work can help identify whether a particular concept is best captured by one model over others.

The relationship between concepts and problems in philosophy is a perennial issue, but it has received a particular attention in the last few decades. One reason for this focus is that many philosophers—including Frank Jackson, David Lewis, and David Chalmers—have argued that traction can be gained in solving philosophical problems if we identify the structure of broadly shared concepts. The idea is that if the concepts of rational, well-informed people tend to be more or less the same on a given philosophical issue, then their judgments will reflect principles, platitudes, or well-hewn truths that are also prescriptive. This particular view of conceptual analysis is not the only game in town; revisionists concerning free will think that our intuitions should be resisted or rejected in the face of certain considerations about well-known biases or principles. Manuel Vargas is a representative of this position. But revisionism remains a minority view.

Robert Kane's libertarian account of free will is not revisionist—or, at least, it's not intended to be so. Kane frequently begins explaining his account by reaching back into history to cite important moments in free will debates. When he does this, he is interested in seeing what is important to them; this often means asking, what did

1. In other words, a concept of free will is a subset of the knowledge about free will that we store in long-term memory, namely, the part of our knowledge that is used to categorize a particular action as a free action or as a not-freely-performed action (Belohlavek and Klir 2011; Machery 2009).

philosophers think best characterizes what people wanted free will to do? Why do many people appear to think that free will is necessary for morality and practical reason? And what metaphysical conditions are requisite for the possibility of such a status? Kane then attempts to synthesize this information without loss, asking the question: what now must the world be like in order for this concept to be correctly applied? In addition, Kane has denied that he is even a moderate revisionist who would endorse even "pruning" our everyday concept (Vargas 2005). So I think we should interpret Kane's account as he intends, and hold that his view is meant to capture the ways ordinary people generally think about free will.[2]

And in Kane's view, nonphilosophers are incompatibilists, at least until philosophers come along and convince them otherwise. On this view, the man on the street believes determinism and free will, in the sense necessary for moral responsibility, are incompatible. Kane writes,

> In my experience, most ordinary persons start out as natural incompatibilists. They believe there is some kind of conflict between freedom and determinism; and the idea that freedom and responsibility might be compatible with determinism looks to them at first like a 'quagmire of evasion' (William James) or 'a wretched subterfuge' (Immanuel Kant). Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers. (Kane 1999, 217).

Kane is not alone in this opinion about what description best characterizes people's general concept of free will. Many philosophers have proposed that nonphilosophers are best described as natural incompatibilists. Galen Strawson, for instance, writes that incompatibilism describes "just the kind of freedom that most people ordinarily and unreflectively suppose themselves to possess" (Strawson 1986, 30). Similarly, Derk Pereboom writes, "Beginning students typically recoil at the compatibilist response to the problem of moral responsibility" (Pereboom 2001, xvi).

But in addition to holding that people are natural incompatibilists, Kane supposes the concept of free will to be organized in such a way that it could be given sufficient conditions that may be used across a range of cases. This presupposes that

2. Kane is aware that there are many ways in which people think about free will; my target is at least one common among them that is a) "a significant kind of freedom worth wanting" and b) that is incompatible with determinism (Kane 1998, 14–15).

nonphilosophers—except where they are in error—will betray a cognitive model or concept of free will, which conforms to the structure of a definition. If one is judging correctly according to this concept, and all the necessary components of free will are present in a given situation, we should expect this person to conclude that the action was freely chosen and otherwise that the action was not free.

With this metaphilosophical model at work, Kane presents a two-stage theory that is now quite well-known. Two-stage models of free will posit that we have, at the first stage, a capacity for first generating considerations in a nondeterministic way and then, at a second stage, choosing among considerations according to the determination of our will. Such models purport to explain how agents in a given circumstance can have multiple considerations available to them—thus allowing for randomness in the possible paths an action can take—while also accounting for the fact that their decision is brought about by a sufficiently determined will.

Kane cites dissatisfaction with most two-stage theories, which he feels do not go far enough in adequately accounting for the concept of free will. He instead proposes a two-stage model that allows for "dual rational control" or, the ability to do otherwise in precisely the same circumstance (Kane 1985). This involves inserting indeterminacy not only in the moment of the accumulation of alternative considerations, but also in the determination of the decision itself. Although not all decisions will have this indeterminacy and many of our actions will flow directly from a determined character, Kane argues that it is necessary for free will that, at some point in the past, free agents engaged in self-forming actions. In *The Significance of Free Will*, his examples of a businesswoman who decides between self-interest and conscience, and an engineer who decides between his craving for alcohol and his desire to save his marriage and career are intended as examples of such actions. Kane describes these cases as involving "recurrent and connected neural networks" for both sides of the issue, both of which are reflections of the character of the individual. As these networks run their course, a chaotic and amplifying interaction is produced where both networks interfere with each other as they run toward the output of a decision. The result is that "the uncertainty and inner tension that agents feel at such moments are reflected in the indeterminacy of their neural processes" (Kane 1998, 130). Kane believes such a model entails that we may later be responsible for our actions, even where no immediate alternative possibilities exist, provided that our action proceeds from our character, where our character is the product of undetermined self-forming actions.

Kane then argues that the free will debate has been mired in equivocation, since incompatibilists but not compatibilists are committed to ultimate responsibility, a desire

to be the ultimate creators of value and the sources of our own nature (Kane 1996, 58–78). The important question is then whether incompatibilism—particularly the branch committed to a need for ultimate responsibility—best characterizes the substantive issue of free will.

II. Are Nonphilosophers Compatibilist or Incompatibilist?

The short answer is that while the body of evidence has largely suggested that compatibalism best characterizes pre-theoretical intuitions, there is presently no consensus concerning whether incompatibilism or compatibalism best characterizes the ordinary concept of freedom of the will. The reason for this is that different ways of asking about free will provoke different responses. Eddy Nahmias has provided evidence that ordinary intuitions about free will are compatibilist. Nahmias et al. (2006) presented participants with a variety of different scenarios describing deterministic universes and for each scenario, participants were asked whether a person in that scenario acted freely and could be held morally responsible for their action. One scenario described a universe capable of being exactly predicted at any given moment by a supercomputer. In this scenario, a man robs a bank and the participant is asked if the man is morally blameworthy for his action. Most participants (76%), when presented with this information, say yes. In other words, despite the computer's ability to predict exactly what the man would do, most participants thought that he is still responsible for his action. In follow-up questions, roughly two-thirds of participants claimed that agents in these worlds had free will and slightly more than four-fifths claimed that agents have moral responsibility.

Nahmias takes this to be strong evidence for a compatibilist description of pre-philosophical intuitions. But does characterizing determinism as "a capacity to provide exact prediction of human action" capture what philosophers mean by determinism? If it does, it does so indirectly. Nahmias and colleagues are taking this way of operationalizing free will from Sam Harris's book, *Free Will*, which employs a scenario like this to motivate the idea that free will is an illusion. But Peter Van Inwagen was closer to the truth when he wrote that determinism, in its most basic formulation, holds that "there is at any instant exactly one physically possible future" (Van Inwagen 1983, 3). Perhaps determinism necessarily entails that a supercomputer could, in principle, have the ability to predict all human action, but the converse doesn't hold true: in other words, it doesn't follow that because a supercomputer can predict future human actions that those actions are determined, for perhaps the universe does involve random chance, but not in the relationship between future human action and the predictions of the supercomputer,

which at every instance the computer makes its prediction, fixes a relation between its prediction and some human action. So, one could object that the question Nahmias and colleagues have used to operationalize determinism isn't necessarily getting at determinism, even if it is in the ballpark.

Secondly, libertarians, determinists, and compatibilists among us have friends and spouses who are pretty good at predicting our own behavior. Imagining a superlative capability to predict our actions might just be an imaginative extension of this rather mundane fact for the participants in the study. And since those close to us can often predict what we will say and do without detracting from our capacity to be responsible for those things, why should it follow that an exact prediction should do so?

This hunch, that asking about perfect prediction instead of causation might make a real difference, has already been tested. A paper by Luke Misenheimer (2008) tested whether asking about *perfect prediction* or *complete causation* made a difference in the responses. Subjects were presented with a description of an imaginary world in which people have either, in the causation condition, scientists who have discovered that every action of these beings is caused by things that happened before them, or, in the prediction condition, scientists who have discovered that every action of these beings can be perfectly predicted. Subjects were then asked about an individual being in these worlds who embezzles money, and whether or not they could have done so freely. Misenheimer found that whereas 30% of responses signaled agreement in the causation condition—indicating incompatibilism, 63% of responses signaled agreement in the prediction condition, indicating compatibalism.

This is a large effect, and it clearly supports the hypothesis that language that describes causation explicitly does a much better job at securing incompatibilist responses than does predictive language, which returns compatibilist responses. Why would this be the case? One hypothesis is that causation descriptions are mechanistic and defy what Daniel Dennett has called the intentional stance. Normally, when we deal with agents, we attempt to figure out what mental states they might have, given their environment and their behavior. In particular, we attempt to describe what the agent's desires and objectives are given this behavior. We then attempt to predict what an agent will do on the basis of this analysis. But language involving causation drops to a lower level of abstraction and asks us not to take up the position of evaluating the mental states of an agent, but rather to explain their behavior by reference to something else (e.g., brain states).

In an influential paper, Shaun Nichols and Joshua Knobe (2007) also complicated Nahmias's conclusion. They demonstrated that two ways of presenting the question

of determinism had a considerable effect on whether the response was compatibilist or incompatibilist. They first described two universes: Universe A, in which every decision "is completely caused by what happened before the decision—given the past, each decision has to happen the way that it does," and Universe B in which "decisions are not completely caused by the past, and each human decision does not have to happen the way it does" (5). In two additional conditions, they described an *abstract* case, in which they solicited an answer to the question: "In Universe A, is it possible for a person to be fully morally responsible for their actions?," whereas in a *concrete* case they ask whether or not Bill, a man who killed his wife and children in a fire to sleep with his secretary, is morally responsible for the killing.

Nichols and Knobe first asked directly: which universe, A or B, is more like our own? Participants overwhelming choose B, the indeterminist universe. They then asked if, in the concrete condition, Bill was morally responsible for the killing of his wife in Universe A, to which 72% of participants said yes—a finding that confirms the previous findings of Nahmias and colleagues. However, in the abstract condition, 86% of participants answered that a person could not be responsible in Universe A.

This finding has held up, even across cultures. Sarkissian et al. (2010) reports finding that in four distinctive cultures there was a consensus among respondents for two theses: a) that we live in an undetermined universe and b) that moral responsibility is not compatible with determinism. For their study, Sarkissian and his colleagues used a total of 231 undergraduate students from the United States, Hong Kong, India, and Columbia, each group divided roughly in half by sex. They were given the descriptions of universes A and B and asked the same questions. Among the four cultural groups, it was found that there was little difference. This is highly surprising. Researchers have found evidence that culture has a large effect on notions of moral responsibility (Miller & Turnbull 1986), what it means to be an individual (Markus & Kitayama 1991), and even what kinds of fallacies we fall for (Nisbett 2003). But Sarkissian et al. (2010) supports the view that whatever psychological mechanisms are guiding the distinctive incompatibilist judgments in the abstract cases and the compatibilist judgments in the concrete cases, they do not vary widely from culture to culture.

This dramatic contradiction between the abstract and the concrete cases has been a major preoccupation in the experimental work on free will and theories abound. Initially, Nichols and Knobe thought that two distinct psychological processes were driving judgments concerning free will in contrary directions. They hypothesized that in the abstract case, people were moved to say individuals in a causally determined universe did not have free will, but as a case became more concrete, a distinct process leads people to

say a person is, in fact, responsible, whether they live in a causally determined universe or not. Their initial interpretation was that our ordinary judgments are incompatibilist, but that an affective response in the concrete condition leads to performance error.

If this interpretation is correct, then our judgments appear systematically produced in no small part by feeling states. This would give us reason to doubt the two-stage theorist's proposal that our everyday concept of free will can be assessed by a strict consideration of facts regarding agency (e.g., what the agent knew, whether their actions flowed from their character, whether they had alternative possibilities open to them, etc.). Moreover, it would then provide reason to be suspect of Kane's particular description of self-forming actions, whose descriptions involve cases of internal moral conflict and the creation of meaningful contributions to the character of the agent. Such descriptions stir our empathy for an agent, and may contribute to the sympathy or disgust will feel at their decision, and if this hypothesis is correct, these moral sentiments may lead us to attribute an exculpatory or inculpatory status. In other words—to the skeptic's rejoicing—it would appear on this interpretation that Kane's descriptions of self-forming actions are playing to an error theory of free will judgments rather than revealing a deeply latent principle in everyday psychology.

Fortunately for Kane, the evidence has not born out this conclusion and Knobe has since retracted this proposal. The initial conclusion that affect was the driving force behind compatibilist intuitions in the concrete cases has been challenged and largely set aside, despite its initial plausibility. A recent meta-analysis of twenty-nine studies shows that high affect cases do indeed produce compatibilist responses, but that this response is actually quite small (d = .18) (Feltz & Cova 2012).

Nahmias and colleagues have also questioned Knobe's original study for its language. They have found evidence that when the wording that describes determinism is phrased in mechanical or reductionist terms, this can lead to a *bypassing error*, whereby participants assume that the person in the scenario can not be acting according to their own motivations. Nahmias and Murray (2010) presented participants with the following statement: "In Universe A, what a person believes has no effect on what he or she ends up being caused to do." Surprisingly, they found a majority of participants tend to agree. In other words, participants likely infer that causal determination prohibits individuals from acting according to their beliefs and desires. Nahmias has argued that this may be one reason why the free will debate is so bedeviling, namely, incompatibilists will frequently give descriptions of determinism that present it alongside a mechanical or materially reductionist picture, and this description occasions the bypassing error and gives the appearance of a strong incompatibilist response. Likewise, the incompatibilism

that Knobe and Nichols' original study may have found could be the result of bypassing due to the wording of the vignette, rather than a reflection of a more general cognitive process reflecting typical judgments about determinism.[3]

More recently, Knobe and Nahmias have offered two opposing theories to explain the available evidence to date (Machery & O'Neill 2014, 69). Whether either is correct, time will tell. But the foregoing discussion here should be sufficient to highlight the pitfalls of assigning nonphilosopher's to one camp in this debate or another on the basis of anecdotal evidence. And while there are now repeated findings in the experimental literature, we should pause to note that what is of interest to philosophers doing this experimental work is not merely what label to attach to survey responses. Rather, they are using experimental methods in partnership with traditional philosophy to both test and generate hypotheses about the underlying structures of the concepts involved. The debate concerning free will is genuinely complex and an easy answer is likely to be elusive. It is likely that there will be some canonical descriptions of determinism that will not bear out certain responses, and other canonical descriptions that will.

III. Alternative Concept Models

I turn now to a second part of Kane's claim: that the ordinary concept of free will is best characterized by a series of necessary conditions. Such a concept model is often called a definitional model, or the classic model of concepts. A frequent example philosophers invoke is the concept of a bachelor, which includes two components: male and single. When these two components are present, then there is a categorization under bachelor and when one or more is absent, then there is no categorization under the heading bachelor. Similarly, Kane's account of free will holds that barring error, we should expect people to assert under some conditions that an action is free and that an agent is morally responsible, and to withhold this judgment when those conditions are not present. How much of the specifics of Kane's account are required to be in these conditions is, I think, up for debate. However, it appears Kane must hold that at a minimum, these conditions should include that an agent has alternative possibilities present at the time of action or previously, and that an agent has a capacity to act according to what they deem to be their own best reasons. When these two conditions are present—perhaps with some addition—then we should expect to see a majority of respondents declaring that an

3. However, evidence of bypassing has been recently challenged by Gunnar Björnsson and Derk Pereboom (2014), who provide evidence that the wording of Nahmias's studies can be understood in ways that involve *passing through* rather than *bypassing* the agent's decision, desires, and beliefs.

agent acted freely and can be morally responsible for an act, and when one or more of these conditions is absent, this declaration should not be found.

In an original study, Josh May has found evidence for a different concept model characterizing free will. May (2013) provides experimental evidence motivating a cluster theory of free will, according to which components of the concept work as features of free will rather than as necessary conditions. These findings support a challenge to traditional philosophical accounts of free will that frequently assume that people work from tacit definitions of free will—definitions that philosophical accounts of free will are often depicted as illuminating rather than revising.

In his study, May tested two aspects of free will that have been critical in the literature: ensurance and liberty. Ensurance is that feature of free will that is marked by the agent's control over their actions. One can think of ensurance as the capacity to act based on her own mental states. Liberty is marked by having alternative possibilities, where an agent has the power to make a genuine choice between more than one option. May hypothesized that non-philosophers would be highly moved to respond that a person acted freely if she had both liberty and ensurance. In cases where both factors were missing, there would be little assertion that she had free will, but in cases where only one factor was present, May predicted that judgments would be mixed.

Using a factorial design, May employed four vignettes and randomly assigned non-philosopher participants. Each vignette featured two paragraphs, where each described a universe that is re-created over and over again from some initial conditions. In the first paragraph, either liberty was present or it was not. When liberty was present, the scenario holds that the laws of nature "needn't" cause the exact same events to happen again, but when liberty was absence, the laws of nature "must" cause the exact same events to happen again. In a second paragraph, a subject named Jill is described as either, in the ensurance condition, "deliberating and deciding" to steal a necklace or, in the no ensurance condition, being "brainwashed to have a powerful urge" to steal a necklace. Each participant in the study was then asked whether "Jill stole the necklace freely"[4] (May 2014, 10).

As expected, May found that the mutual presence of ensurance and liberty encourages judgments that Jill acted freely,[5] their mutual absence finds a nearly complete

4. Responses were recorded using a 7-pt Likert scale, with "disagree completely" and "agree completely" at 1 and 7 respectively and a middle value of "in-between."

5. $\bar{x} = 6.6$; $\sigma = .89$; total in agreement: 96%.

absence of judgment that Jill acted freely,[6] but their individual presence finds mixed results and a lack of consensus.[7] His findings support the hypothesis that the concept of free will has a cluster structure, whose features motivate judgment independently of one another.

Definition models of the ordinary concept of free will can't explain why the presence of ensurance or liberty, in the absence of the other, would result in sizeable minorities of participants responding that an agent had acted freely. The response that this large minority of participants (49% in one case) were simply in error in the case where one feature is present but not the other is unsatisfactory, for the definition model holder can't say this large minority is in error while simultaneously wishing to argue that they are merely clarifying a position rooted in ordinary cognition about free will. The usual response from a two-stage theorist is that a case like the brainwashed Jill case is simply not a case of free will because Jill's actions are not rightfully hers due to a disruption in the flow of actions from her character, her motivations, her reasons, or her desires and beliefs. What this response doesn't account for, given this data, is the fact that this disruption leaves participants without consensus about the free will of an agent, rather than a consensus for the lack of free will of an agent.

These results motivate a cluster theory, but it isn't yet clear from this study which concept model best describes judgments concerning free will. What May has found evidence of are typicality effects in our judgments concerning free will, according to which purported cases of action are not either free or not free, but are rather more or less clearly free according to how much they share certain features common to our concept. Much of the research on typicality and concepts began in earnest in the 1970s with the work of Eleanor Rosch, who found evidence that categories are graded. Both penguins and robins, for instance, are birds, but a robin is more like a bird than a penguin. This gradation also allows psychologists to explain how *systematically* inconsistent people can be in their reasoning. For instance, participants in a study may assent that a dentist's chair is a chair and that chairs are instances of furniture while denying that a dentist chair is a piece of furniture. This intransitivity would lead one to think that people were simply in error about furniture if it was thought that people really had concepts that took the form of definitions, but if we think of the concepts for each of these items as clusters of features, then we can see that they have overlapping characteristics in some areas that

6. $\bar{x}$=3.17; σ =2.25; total in agreement: 30%.

7. No ensurance, but liberty: $\bar{x}$=3.98; σ=2.25; total in agreement: 39% and no liberty, but ensurance: $\bar{x}$=4.67; σ=2.32; total in agreement: 49%.

do not overlap in others, so this inconsistent triad is actually consistent at the level of conceptualization.

Two important models for thinking about cluster theories include exemplars and prototypes. Exemplar models hold that when we categorize an instance under a concept, we look in long-term memory at a collection of the best possible instances of that concept. We then measure the similarity of a given instance to the set of these examples and if a threshold for similarity is achieved, then that instance is categorized under the concept. Prototypes, on the other hand, are best thought of as collections of property data rather than as mental representations of exemplars. James Hampton's formal prototype model is a useful for thinking about prototypes (Hampton 1993). May does not cite Hampton's model but his thinking about the issue is captured by this formalism quite well. Hampton's model includes a similarity measure and a decision rule. The similarity measure, S (x, C), of an instance x to a category C is defined in terms of values w(x, i), which includes the weight of the value (e.g., ensurance) possessed by x for attribute I of the prototype (e.g., free will).

$$S\,(x, C) = \sum_{i} w(x, i)$$

This measurement of similarity can then be used in a decision rule, which states that if the similarity measure (S) of an instance (x) to a category (C) is greater than some threshold (t) then that instance is a member of the category.

$$S\,(x, C) > t \Rightarrow x \in C$$

In coming into contact with objects in the world, we extract from them statistical information concerning how regular a class of properties are associated with objects of a kind. Birds, for instance, can often fly and do so in groups. Birds typically dive into water and build nests; they have beaks, feathers, and forward-facing eyes and are often small. This property description then gets an ordering based on weights. Feathers and beaks have a heavier weight, whereas flying in groups and relative size have lighter weights. This allows us to account for how humming birds and ostriches can both be birds, while being less typical birds than common songbirds.

This way of approaching the question of free will makes it clear that May's study is really only the beginning of a promising empirical approach. Future studies should look not only at ensurance and liberty as attributes whose weights are measured, but also at the many components philosophers have long thought—sometimes controversially—to be essential to exercising free will. In addition, May does not discuss the thesis of Edouard

Machery's book, *Doing without Concepts* (2009). Machery proposes a heterogeneity thesis concerning concepts, which holds that most categories (e.g., free actions) are represented by several concepts that belong to kinds that have little in common. Each of these kinds, which may include prototypes, exemplars, and theories, can be involved in a variety of cognitive processes, including learning, recalling, revising, induction, and judgments. If this thesis is true, it implies that the concept of free will may well consist of distinct bodies of knowledge rather than as a single body of knowledge. This also means it is possible that future experimental philosophers will discover that Kane's significant conception of free will does reflect some bodies of knowledge regarding free will, while failing to adequately capture others.

Conclusion

I have argued that presently available studies on the shared concept of free will do not provide clear empirical support that people are generally incompatibilist; rather, different cognitive processes—reflected in different ways of setting up individual questions—produce different results. Whether or not some of these different ways of asking questions are relevant is a question for further reflection and argument. I have also argued that there may be reason—however under-supported—to doubt the intuitive force of Kane's notion of self-forming actions, due to the affect such descriptions produce, and the role such affect might have in over-riding our usual responses. Finally, I have argued that May provides compelling evidence that the structure of nonphilosopher's concept of free will is distinct from the structure described by Kane. A similar argument has been made now by many others, including Stephen Stitch, Alvin Goldman, and Mark Johnson, who have each argued that the demise of the definitional model should give us reason to reject normative ethical theories that presuppose nonphilosophers to represent normative categories using necessary and sufficient conditions. Further evidence may continue to demonstrate a similar argument for philosophical accounts of action that presuppose a definitional model.

References

Beebe, James, and Wesley Buckwalter. 2010. "The epistemic side-effect effect." *Mind & Language* 25 (4): 474–498.

Belohlavek, Radim, and George J. Klir. 2011. *Concepts and Fuzzy Logic*. Cambridge: The MIT Press.

Björnsson, Gunnar, and Derk Pereboom. 2014. "Free Will Skepticism and Bypassing." *Moral Psychology*, Volume 4, edited by Walter Sinnott-Armstrong, 27–36. Cambridge: The MIT Press.

Feltz, Adam, and Florian Cova. 2012. "When and how affective reactions impact judgments about free will and determinism: A Meta-analysis." Unpublished manuscript.

Kane, Robert. 1985. *Free Will and Values*. New York: SUNY Press.

Kane, Robert. 1998. *The Significance of Free Will*. New York: Oxford University Press.

Kane, Robert. 1999. "Responsibility, luck, and chance: reflections on free will and indeterminism." *Journal of Philosophy* 96 (5): 217–240.

Hampton, James A. 1979. "Polymorphous concepts in semantic memory." *Journal of Verbal Learning and Verbal Behavior* 18 (4): 441–461.

Hampton, James A. 1996. "Testing the prototype theory of concepts." *Journal of Memory and Language* 34: 686–708.

Machery, Edouard. 2009. *Doing without Concepts*. New York: Oxford University Press.

Machery, Edouard, and Elizabeth O'Neill. 2014. *Current Controversies in Experimental Philosophy*. London: Routledge.

May, Joshua. 2014. "On the very concept of free will." *Synthese* 191 (12): 2849–2866.

Markus, Hazel, and Shinobu Kitayama. 1991. "Culture and the self: implications for cognition, emotion, and motivation." *Psychological Review* 98 (2): 224–253.

Miller, Dale T., and William Turnbull. 1986. "Expectancies and interpersonal processes." *Annual Review of Psychology* 37 (7): 233–256.

Misenheimer, Luke. 2008. "Predictability, causation, and free will." Accessed 6/10/14: http://philosophy.berkeley.edu/file/551/misenheimer-free_will.pdf.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2006. "Is incompatibilism intuitive?" *Philosophy and Phenomenological Research*. LXXIII (1): 28–53.

Nahmias, Eddy, Justin Coates, and Trevor Kvaran. 2007. "Free will, moral responsibility and mechanism: Experiments on folk intuitions." *Midwest Studies in Philosophy* XXXI (1): 214–242.

Nisbett, Richard. 2003. *The Geography of Thought*. New York: Free Press.

Park, John Jung. 2011. "Prototypes, exemplars, and theoretical & applied ethics." *Neuroethics* 6 (2): 237–247.

Pereboom, Derk. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Rosch, Eleanor. 1975. "Cognitive representations of semantic categories." *Journal of Experimental Psychology: General* 104 (3): 192–233.

Sarkissian, Hagop, et al. 2010. "Is belief in free will a cultural universal?" *Mind & Language* 25 (3): 346–358.

Strawson, Galen. 2010. *Freedom and Belief*. New York: Oxford University Press.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, Manuel. 2005. "The revisionist's guide to responsibility." *Philosophical Studies* 125 (3): 399–429.

Journal of Cognition and Neuroethics

The Illusion of Freedom: Agent-Causation and Self-Deception

Jacob Quick
Northern Illinois University

Biography

Jacob Quick is completing a Master of Arts in Philosophy at Northern Illinois University, where he has focused on epistemology and political philosophy. Jacob's other areas of interest include philosophy of action, philosophy of religion, and continental philosophy.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Quick, Jacob. 2015. "The Illusion of Freedom: Agent-Causation and Self-Deception." *Journal of Cognition and Neuroethics* 3 (1): 297–308.

The Illusion of Freedom: Agent-Causation and Self-Deception

Jacob Quick

Abstract
My purpose in this paper is to argue in favor of the external observer and show that Campbell is not justified in merely relying on the testimony of the acting agent. First, I will present and explain the main tenets of Campbell's libertarian agent-causation. Second, I will analyze Campbell's defense of agent-causation. Third, I will present data gathered from psychological studies suggesting that acting agents are mostly unaware of the factors which comprise their actions. Fourth, I will present recent work done on the psychology of self-deception and how this research discredits the testimony of the acting agent. Finally, I will summarize my argument and discuss the implications of my argument for Campbell's motivation for agent-causation.

Keywords
Agent-causastion, self-deception, introspection, epistemic privilege, psychology, Campbell, acting agent, external observer

In C.A. Campbell's *In Defence of Free Will*, Campbell defends an agent-causal theory of free will on the basis that a subject experiences himself as the uncaused cause of morally significant actions. However, the 'external observer' interprets another agent's actions as determined by causal antecedents apart from the acting agent. Thus, when S performs a morally significant action P, S interprets S as the *sole* cause of P. However, when an external observer T examines P, T interprets P as determined, at least in part, by causal antecedents apart from S. I will refer to S as the 'acting agent' and to T as the 'external observer.' Campbell then argues that the interpretation of the acting agent should take priority over the interpretation of the external observer in the free will debate. He places the burden of proof on the opposing side and bemoans the lack of literature that determinists have provided in favor of the external observer (Campbell 1967, 50).

Campbell's argument still plays a role in current discussions of free will, specifically concerning agent-causal theories. Campbell's thesis brings up some important questions: should priority be given to the interpretation of the acting agent or that of the external observer in the free will debate? When the results of external observation seem to conflict with our intuitions and beliefs concerning our own free actions, should we give priority to our intuitions or to the results of our observations? Recent psychological experiments

which show that agents are highly prone to self-deception and faulty self-assessment comes to bear upon analyzing these questions.

My purpose in this paper is to argue in favor of the external observer and show that Campbell is not justified in merely relying on the testimony of the acting agent. First, I will present and explain the main tenets of Campbell's libertarian agent-causation. Second, I will analyze Campbell's defense of agent-causation. Third, I will present data gathered from psychological studies suggesting that acting agents are mostly unaware of the factors which comprise their actions. Fourth, I will present recent work done on the psychology of self-deception and how this research discredits the testimony of the acting agent. Finally, I will summarize my argument and discuss the implications of my argument for Campbell's motivation for agent-causation.[1]

I. Agent-Causation

Now I will present and explain the main tenets of Campbell's libertarian Agent-Causal view (AC). AC is an indeterministic view. Thus, AC maintains that agents have free will and that the free will that agents possess is incompatible with determinism. According to AC, when an agent performs a free action in a specific situation, that agent could have performed a different action in that exact situation, at the same time, and given the same past. In AC, the agent's free action is not caused by anything other than the agent. Neither reasons, nor desires, nor a state of affairs can produce the free action of an agent. The agent cannot be an effect of a prior cause. As a result, the agent must *solely* bring about a particular, free action (Campbell 1967, 43).[2]

II. Motivations for AC

Campbell admits that the metaphysics of AC can be complicated and confusing. Campbell also explains that he is not motivated to hold to AC on the basis of any conceptual clarity. Rather, Campbell proposes that AC is attractive because it coheres with the perspective of the acting agent (AA). As I have noted earlier Campbell maintains

1. It is important to note that my intention is not to completely discredit the testimony of the acting agent. Rather, I want to show that the acting agent does not have the kind of epistemic privilege required in order for Campbell's defense of agent-causation to be successful.

2. It should be noted that not all proponents of agent-causation maintain all of the tenets just listed. In fact, Randolph Clarke presents a different, less radical account of agent-causation (Clarke 1993, 191-203). However, since my argument specifically focuses on Campbell's defense, I will confine my discussion of AC to Campbell's account.

that when S performs a morally significant action P, S interprets S as the *sole* cause of P. However, when an external observer T examines P, T interprets P as determined, at least in part, by causal antecedents apart from S. In Campbell's account S is the acting agent (AA) and T is the external observer (EO). Campbell asks why humans believe that they are uncaused causes of their moral actions, and provides what he believes to be the best answer: "They do so, at bottom, because they feel certain of the existence of such activity from their immediate practical experience of themselves" (1967, 41). Campbell seeks to explain that it is in the situation of moral temptation that we experience our actions as originating solely within the self apart from desire, heredity, etc.

Campbell holds that the unintelligibility objection to AC only succeeds if one takes the position of EO. However, Campbell argues, the proper standpoint to take concerning free acts is that of AA. Campbell argues that it is an error for one to examine and discern the nature of free moral actions from the perspective of EO:

> It is perfectly true that the standpoint of the external observer, which we are obliged to adopt in dealing with physical processes, does not furnish us with even a glimmering of a notion of what can be meant by an entity which acts causally and yet not through any of the determinate features of its character. So far as we confine ourselves to external observation, I agree that this notion must seem to us pure nonsense. But then we are not obliged to confine ourselves to external observation in dealing with the human agent. Here, though here alone, we have the inestimable advantage of being able to apprehend operations from the inside, from the standpoint of living experience. But if we do adopt this internal standpoint – surely a proper standpoint, and one which we should be only too glad to adopt if we could in the case of other entities – the situation is entirely changed. We find that we not merely can, but constantly do, attach meaning to a causation which is the self's causation but is not yet exercised by the self's character. (Campbell 1967, 48)

Thus, Campbell concedes that, from the standpoint of EO, AC is a nonsensical notion. However, AC accurately describes how things appear from the standpoint of AA and, moreover, the interpretation of AA should be given more weight when discerning the nature of free, morally significant actions. Thus, Campbell posits that AA should have *epistemic privilege* over EO concerning the precise nature of free moral actions. My definition of epistemic privilege, for the purpose of this paper, is as follows:

> **Epistemic Privilege**: S has epistemic privilege if and only if S's interpretation concerning a certain subject P is presumed to most accurately correspond to the actual nature of P.

Campbell posits that if it were the case that an agent's causing a certain action could happen without relation to the acting agent's character, then the *only* way in which we could be aware of such a thing is from the perspective of AA. Campbell asserts that the only legitimate way in which one could criticize his position is to present "a reasoned justification of his cavalier attitude towards the testimony of practical self-consciousness. That is the primary desideratum" (1967, 50). My aim in this paper is to provide the very justification against the epistemic privilege of AA that Campbell demands.

While there is more literature critiquing the epistemic privilege of AA now than there was during Campbell's time, there are still proponents of agent-causation who find Campbell's motivation for AC compelling. For instance, Timothy O'Connor, perhaps the most prominent contemporary defender of an agent-causal account, contends:

> …the agency theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favor doings so; it seems to be the case, rather that I produce my decisions in view of those reasons, and could have, in an unconditional sense, decided differently… Such experiences could, of course, be wholly illusory, but do we not properly assume, in the absence of strong countervailing reasons, that things are pretty much the way they appear to us? (O'Connor 1995, 196).

Thus, Campbell's motivation for AC is still utilized in the free will discussion.

There are multiple ways to approach Campbell's argument. For instance, Mele argues that, contrary to the claims of Campbell and O'Connor, AA does not actually experience his own free actions as agent-caused (Mele 1995). However, my aim is not to contend with whether AA does or does not interpret her own experience as agent-caused. Rather, my contention is whether the perspective of AA can justifiably be utilized as a strong and persuasive argument on behalf of AC. Campbell challenges opponents of AC to provide data that disputes the idea that AA has epistemic privilege. In what follows, I will provide data gathered from multiple psychological experiments that discredits the epistemic privilege of AA.

III. Psychological Data

Richard Nisbett and Timothy Wilson performed psychological experiments that displayed the propensity of agents to be unaware of environmental influences upon their motivations and judgments.[3] I will present two of their experiments and note the conclusions drawn from these experiments.

Nylon Stockings Experiment

In the Nylon Stockings experiments four identical nylon stockings were placed in a row. Participants were asked to judge the quality of the stockings and discern which stocking was superior to the others. The experiment was designed so that the subjects would examine the leftmost stocking first and going down the row, end the inspection by examining the rightmost stocking. The left-to-right positioning of the stockings had a major effect on the subjects' judgments. In fact, subjects were almost four times more likely to prefer the right-most stocking over the left-most stocking. Nisbett and Wilson note the response that participants gave when it was suggested that the positioning of the stockings played a role in determining their preferences:

> When asked about the reason for their choices, no subject ever mentioned spontaneously the position of the article in the array. And, when asked directly about a possible effect of the position of the article, virtually all subjects denied it, usually with a worried glance at the interview suggesting that they felt either that they had misunderstood the question or were dealing with a madman. (Nisbett and Wilson, 1977, 243-244)

The Nylon Stockings experiment was repeated by using nightgowns instead of stockings. The left-to-right positioning played a major role in the subjects' choice of nightgown and confirmed the results of the Nylon Stockings Experiment (Nisbett and Wilson 1977, 243).

The European Professor

In another experiment, subjects were shown a video of a college teacher, who spoke English with a European accent, responding to a student's question. After watching the video, the subjects were asked to rate their appreciation of the teacher and their

3. I would like to thank Dr. Neil Otte for bringing the work of Nisbett and Wilson to my attention.

appreciation of the teacher's appearance, accent, and mannerisms. Half of the subjects saw the teacher answering the student's question in a warm, agreeable manner, while the other subjects saw the professor answer coldly. However, in both videos, the teacher's accent, mannerisms, and appearance remained the same. Those who saw the professor answer warmly rated the teacher's accent, mannerisms and appearance as attractive, while the majority of participants who saw him answer coldly found the teacher's qualities to be irritating. Nisbett and Wilson note that participants in both groups were asked whether their ratings of the teacher's qualities were affected by their appreciation of the teacher. Likewise, participants from both groups were asked whether their appreciation for the teacher's attributes affected their appreciation of the teacher. The participants in both warm and cold groups denied any causal connection between their impression of the teacher and their impression of his attributes. Also, all of the subjects in the warm group, who were asked, denied that their appreciation of the teacher's qualities affected their appreciation of the teacher overall. However, some of the participants in the cold version reported that their dislike of the teacher's qualities lowered their overall appreciation of him. Thus, the participants denied what was actually happening (their overall appreciation of the teacher affected their appreciation of his qualities) and some even inverted the causal relationship (Nisbett and Wilson 1977, 244-245).

The data that Nisbett and Wilson present suggests that we can commonly misunderstand the nature of our motivations, judgments, and interactions. In their experiments, AA is not aware of the effect that the external environment has on her acting states. While I only cited two experiments, Nisbett and Wilson utilize multiple experiments that suggest that we are not reliable informants concerning the nature of our own choices and actions. In fact, after examining and conducting their experiments, Nisbett and Wilson conclude: "The accuracy of subjective reports is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports" (1977, 233). Thus, the experiments strongly suggest that the AA is not reliable and, therefore, does not have the Epistemic Privilege that Campbell's view requires.

IV. Self-Deception

Now, I will argue that recent studies on self-deception show that the perspective of AA should not be considered to have epistemic privilege concerning the nature of free action. There is a great debate, particularly in philosophical circles, over the nature and

existence of self-deception. My goal is not to recount the specifics of the debate.[4] Rather, a great deal of psychological literature utilizes some notion of self-deception and appears to have found strong evidence in favor of it. Thus, for the purposes of this paper, I will utilize a notion of self-deception that accords with the phenomenon that continually arises in psychological experiments and studies. I find that Mele's notion of self-deception best articulates the phenomenon that psychologists find without falling into bewildering paradoxes.[5]

Mele's account of self-deception is strongly associated with desire. According to Mele, we often describe someone as self-deceived because they believe something that they want to believe, even though there is significant evidence to the contrary. Certain forms of ignoring evidence, biased interpretations, and etcetera, lead to self-deception. Mele offers a set of sufficient conditions that accurately describe S entering into self-deception:

> (i) The belief that *p* which S acquires if false.
>
> (ii) S's desiring that *p* leads S to manipulate (i.e., to treat inappropriately) a datum or data relevant, or at least seemingly relevant, to the truth value of *p*.
>
> (iii) This manipulation is a cause of S's acquiring the belief that *p*.
>
> (iv) If, in the causal chain between desire and manipulation or in that between manipulation and belief-acquisition, there are any accidental intermediaries (links), or intermediaries intentionally introduced by another agent, these intermediaries do not make S (significantly) less responsible for acquiring the belief that *p* than he would otherwise have been. (Mele 1983, 370)

I find that Mele's account of self-deception accurately describes the characteristics of self-deception discovered in psychological literature while remaining philosophically coherent.

4. Jeffrey Foss provides a helpful and thorough analysis of the various articulations of self-deception (1980, 237-243).

5. Such as the paradox of an agent intentionally deceiving himself into believing a proposition that he knows to be false. The paradoxes of the sort just mentioned can be found in the articulation of self-deception presented by Raphael Demos (1960, 588-595).

Thus, the reader can assume that when I use the term 'self-deception', I am utilizing Mele's articulation.

Experiments in Self-Deception

Robert Trivers and William von Hippel, two psychologists who have done a great amount of research on self-deception, note multiple experiments in which agents deceive themselves *about themselves and their own actions*. Thus, in certain circumstances, agents have a high propensity for believing false information about the nature and details of their own actions. I will briefly present the results of multiple studies focused on self-deception.

Memory

Psychologists Trivers and von Hippel note that an agent's desires and preferences can cause the agent to misremember certain information about themselves and previous performances. In an experiment in which subjects participated in a study skills course, the participants remembered their original study skills, prior to the course, as lower than they actually were. Participants were prone to this deception because they strongly desired for their skills to improve as a result of the course. Likewise, a little while after the course was finished, the participants had to recount their performance upon completing the study skills course. The participants rated their final performance as higher than it actually was. Thus, the subjects' memories about themselves and their own actions were skewed because of their desire to improve. The subjects' desires led to their self-deception in falsely remembering their beginning performance as worse than it was and their final performance as greater than it actually was (Von Hippel and Trivers 2011, 10).

Rationalization

Research on self-deception and rationalization suggests that we often choose to do certain actions that we deem to be false or wrong when we are better able to rationalize our actions. Thus, our deceptive capacity extends to our decisions in situations of moral temptation. In one experiment, individuals that demonstrated a self-serving bias were placed in circumstances in which they had the ability to cheat. In one situation, the cheating was obviously intentional. In the other situation, the cheating was clearly intentional, but was easier to represent as unintentional due to particular factors in the setting. Those who were able to construe their cheating as unintentional committed the act, while those in the more obvious situation did not. Psychologists suggest that this

phenomenon occurs because, when other environmental factors are present, agents have the ability to deceive themselves into misremembering the intentionality of their action and attribute their action to the environmental factors. In the same vein, it has been shown that people who are told that free will is merely an illusion are more likely to cheat due to their ability to attribute their actions to external factors, obviating them of responsibility (Von Hippel and Trivers 2011, 10).

In another experiment, participants entered a room with two televisions and a disabled person sitting in front of one of the televisions. In some of the cases, both televisions were tuned in to the same program. In other cases, the televisions were tuned to different channels. The participants who walked in and saw that the televisions were on the same channel sat next to the disabled person, while the participants in the room with televisions on different channels sat away from the disabled person and in front of the opposite television. Researchers concluded that the participants who chose to sit away from the disabled person did so because they were able to deceive themselves about their action and claim that they did not sit away from the disabled in order to avoid the disabled, but because they wanted to watch the program that the other television was airing.

Self-deception plays a major role in our behavior in relation to people of another race. A study noted that white people were less likely to give aid to black people than to white people, but only when they could blame other environmental factors such as distance or risk. Thus, white people would help other white people whether or not there were obstacles present. However, white people only helped black people when there were no obstacles present. Since race was the variable, the study showed that the white participants have an implicit preference towards persons of their own race. The participants in the experiment utilized the presence of obstacles in order to justify their disregard for the needs of a black person and therefore deceive themselves into not attributing their action to racial preference. In the words of von Hippel and Trivers, "[The participants] are denying the socially undesirable motives that appear to underlie their behaviors by rationalizing their actions as the product of external forces." (2011, 10).

Recent psychological studies and experiments suggest that self-deception is universally prevalent among agents. Self-deception has been connected with survival and success, which suggests that our ability to distort the truth to others and ourselves is an evolutionary design that best equips us for survival and flourishing (Von Hippel and Trivers 2011, 12-13). We are highly prone to deceive ourselves concerning our character, nature, and our morally significant actions. As a result, we are not justified in assuming that the perspective of AA most accurately corresponds to reality. On the contrary, the

perspective of AA is highly prone to distorting the nature of the agent's actions and deceiving the acting agent.

V. Conclusion

Campbell placed the burden of proof on EO to discredit the epistemic privilege of AA. I believe that recent psychological data demonstrates that AA does not have epistemic privilege with regard to the subject of free will. Psychological experiments indicate that our choices and actions can be heavily influenced by factors of which we are unaware. Studies in self-deception also demonstrate that we have a high propensity toward self-deception concerning our desires, our character, and the nature of our own actions. As a result, Campbell is not justified in granting epistemic privilege to AA.

References

Campbell, C. A. 1967. *In Defence of Free Will, with Other Philosophical Essays*. London: Allen & Unwin.

Clarke, Randolph. 1993. "Toward a Credible Agent-Causal Account of Free Will." *Noûs* 27 (2): 191–203.

Demos, Raphael. 1960. "Lying to Oneself." *Journal of Philosophy* 57 (18): 588–595.

Foss, Jeffrey. 1980. "Rethinking Self-Deception." *American Philosophical Quarterly* 17 (3): 237–243.

O'Connor, Timothy. 1995. "Agent Causation." In *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, edited by Timothy O'Connor, 173–200. Oxford: Oxford University Press.

Mele, Alfred. 1995. *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

Mele, Alfred R. 1983. "Self-Deception." *The Philosophical Quarterly* 33 (133): 365–377.

Nisbett, Richard and Timothy Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3): 231–259.

Von Hippel, William and Robert Trivers. 2011. "The Evolution and Psychology of Self-Deception." *Behavioral and Brain Sciences* 34 (1): 1–56.

Journal of Cognition and Neuroethics

Hegel's Concept of the Free Will: Towards a Redefinition of an Old Question

Fernando Huesca Ramón
Meritorious Autonomous University of Puebla (BUAP)
National Autonomous University of Mexico (UNAM)

Biography

Fernando Huesca Ramón has Bachelor studies in Biology and Philosophy from the Meritorious Autonomous University of Puebla (BUAP), Master studies in Philosophy from the National Autonomous University of Mexico (UNAM), and is currently finishing a doctorate project with the subject: "Political Economy in Hegel: Capital and ethical life." His areas of research are Political Economy, Political Philosophy, Aesthetics, Bioethics and the Philosophy of German Idealism. He currently teaches courses on Aesthetics and Modern Philosophy in BUAP and UNAM.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Huesca Ramón, Fernando. 2015. "Hegel's Concept of the Free Will: Towards a Redefinition of an Old Question." *Journal of Cognition and Neuroethics* 3 (1): 309–325.

Hegel's Concept of the Free Will: Towards a Redefinition of an Old Question

Fernando Huesca Ramón

Abstract

The "will" is a subject of reflection *par excellence* throughout Modern Philosophy. Even Schopenhauer and Nietzsche, two stark critics of the thought agenda of the *Modern* philosophers made the "will" a center of reflection. In this paper, we intend, firstly, to tackle the question of the will, centering upon the subject of whether this "will" is free or not, and what is peculiar, on the human being, concerning this "will;" this will take us to two great thinkers of the 17th century, Spinoza and Hobbes, who in one way or the other, challenged the Cartesian notion of a "free will" or a "soul" which can freely command the body to action. We shall briefly, then, touch upon Kant's view on the matter, highlighting his conclusion (from the third antinomy of the *Critique of Pure Reason*), of the undecidibility of the question of the "free will." Secondly, we shall focus our attention on Hegel's own view on the question of the will, in general, and his thoughts on the peculiarity of the human will. In this way we will embark upon an explanation of Hegel's *solution* of the *problem of the free will*, highlighting his redefinition of the problem, and the way he challenges Kant's skeptic stance, in such a way that we, in the present may, even in an empirical experimental way, ascertain the validity of his arguments. In the end, as a way of conclusion we shall, in effect, explicate, how Hegel' concept of the free will can be understood in Neurophysiological terms, and concomitantly be tested experimentally. In this way we intend to defend the notion of Hegel as a key thinker in the question of the will, and as a source of concepts and reflections, to guide philosophical and experimental research on the matter.

Keywords

Freedom, will, idealism, executive function, causality

> Freedom, Sancho, is one of the most precious gifts which were given to men by the heavens; with her no treasure hidden in the earth or sea can be made equal: for freedom, as for honor, we could and should risk our own life.
>
> —Miguel de Cervantes Saavedra

I

In order to explicate Hegel's concept of the *free will* we shall, firstly, sketch his general ontological position, tackling the notions of the *natural*, and the *spiritual*, and secondly, the anthropological implications of the Hegelian monistic ontology. With this

philosophical background in mind, we shall develop the complex Hegelian argument concerning the *human will*, which in the end shall be presented as an alternative to other modern theories of the will, such as Hobbes', Spinoza's and Kant's which deal with the subject *strictly* and *totally* in the "old" terms of "freedom from causality", that is, of the lacking of previous conditions to movement or change in general, being it material or mental; that will lead us to a clear understanding of the German philosopher's *compatibilist* position on the subject of the free will. Hegel's redefinition of this central subject in Philosophy of Mind opens the way for an ample dialogue and interaction between Philosophy and Neurobiology which can scarcely be stated on the basis of previous (and even later) ontological systems, and which will be sketched as a manner of conclusion.

Let us begin with a substantive Hegelian ontological fragment:

> Man is, on the one side, a natural being (*natürliches Wesen*). As such, he conducts himself according to arbitrariness and chance; as a restless, subjective being. He does not distinguish between the essential, and the unessential. On the other side, he is a spiritual, rational being (*geistiges, vernünftiges Wesen*). From this side *he is not from nature, what he should be*. […] Man must bring his two sides in agreement; that is, to make his singularity adequate to his rational side, or to make this one, the dominant one. (Hegel, 1986a, 258)[1]

This passage taken from a philosophical handbook,[2] prepared by Hegel for *Gymnasium*[3] students presents the nucleus of the Hegelian Ontology; in other words, the principle, that man is *on the one side* "a natural being", and *on the other side* "a spiritual, rational being" states a "monistic" worldview, which is neither materialistic *à la* Hobbes, nor

1. All translations in this paper from German and Spanish sources are ours.

2. That is the so called *Pflichtenlehre* (Doctrine of duties): *Rechts-, Pflichten-* und *Religionslehre für die Unterklasse* (Doctrine of right, duties und religion for the elementary class). This bibliographic source can be considered as a summary or didactic exposition, thought for "secondary-school" students, of the Philosophical doctrine of Right presented in the *Philosophy of Right*.

3. *Das Gymnasium* was in Hegel's time an educational institution, whose main aim was to prepare German youths for University "superior" studies, especially those studies concerning the Humanities or what would later be called *Geisteswissenschaften* (Social sciences). Hegel wrote four reports between 1810 and 1816 concerning the goals of "secondary" education and the teaching of Philosophy in Secondary Schools (such as the *Gymnasium*) and University; they are published in volume 4 of the Suhrkamp edition of Hegel's Complete Works (Hegel, 1986a).

pantheistic (in the sense of the thesis: "all is God") *à la* Spinoza; the task to explicate, in detail, all the notions and categories in play in such a philosophical view would take us far from our main task in this work, given that Hegel himself does not defend a radical revolutionary worldview which rejects previous hilosophical positions (as Kant,[4] for instance, does); rather, he intends to question and criticize previous Ontologies, from the Presocratics to Schelling, in order to make evident, the inherent contradictions and partiality of conclusions and results of previous Philosophies while he integrates them in a complex philosophical architectonic. In that sense, the full explication of the Hegelian standpoint would coincide with a wide critic of the History of Philosophy until Schelling. Nevertheless, we do intend to present the *gist* of Hegel's Ontology; for that purpose, let us state the Ontological basis for understanding the anthropological view presented in the passage above: man is an ontological unity, in such a way, that it is not feasible to invoke a supernatural point of view (which Kant does) in order to understand his character as a "material" existence and as an entity capable of "mental operations;" on the contrary, Hegel's proposal is to consider, at all times, man (and any other philosophical subject) as a *unity*, as a convergence of diverse categories or modes of being/thought (any idealist Philosophy comes, in one way or the other to such view of the identity of such instances), in a way, that any dualism (the sensible/the ideal, soul/body, inclination/duty, life/concept, etc.) comes to be just a mode in which apparently contradictory terms, are thought of unilaterally and partially (and as such, *wrongly*). So, it is not that the natural and the spiritual consist of different ontological realms, rather, the natural consists of a mode of being/thought of the spiritual, in such a way, that the *natural mode* of being, can be transformed or converted into the *spiritual mode* of being[5] and vice versa.[6]

Now, we must yet sketch precisely, how Hegel defines such a thing as *nature*, or a "natural being," and *spirit*, or a "spiritual being."

Concerning nature, the author of the *Phenomenology of Spirit* states the following: "On the one side, nature means, the natural being, as we find ourselves constituted on different sides immediately; the immediate side of our being" (Hegel, 1974, 76). Such a

4. We consider the Kantian worldview as a specific form of ontological dualism; that is, Kant, in the end, accepts and states, that there *are* two kinds or sets of things (entities, beings, etc.): *noumena* and *phaenomena*.

5. Examples of this change of *mode* can be *any* economic production (of a commodity, a tool, a machine, etc.), and the work of art. In both instances it is possible to argue, in a Hegelian manner, for the "impression "of *human ends* into matter. One could call this impression of human ends: *spiritualization*.

6. Such as in death.

Philosophical standpoint is based upon an Absolute Idealism[7] which defends the thesis that *only* thought or that which is the product of thought is, strictly speaking *real* – *wirklich* -, or even *is* in a strong or higher sense: "Intelligence (*Intelligenz*) knows, that, that which is *thought* (*gedacht*), *is*; and that that which *is*, only *is*, in so far as it is a thought (*Gedanke*)" (Hegel, 1991, 378). So, according to this view, the question for the ontological status of such beings as planets, threes, jellyfish, etc., lies, not in negating that such beings, *exist* or *are*, in one way or the other, rather, the decisive point lies on the notion, that, on the one side, *we*, as human beings (actually, we ourselves define ourselves as such, in an exercise of *thought*!), find such *empirical* or *perceptual* elements, already as given, we do not recognize them as a creation or production of ours, rather as just *being there* when we come to find them within our daily life, experience, or scientific research; on the other side, we, (again, as human beings) consider (with good reasons, one may advance) them precisely as *unable* to come to the understanding or cognition[8] that they themselves are just *given* entities, which *of themselves*, or *caused by* themselves are unable to make out of themselves something different from that which they *already* are; in Hegel's words: "The formation (*Formierung*) of plants, of animals, consists only in maintaining their natural being, or in that this is modified only a little" (Hegel, 1983a,

7. According to F. Beiser, the doctrine of Absolute Idealism has the following traits: "First, there is a single universal substance in nature, which is the absolute. Second, this absolute consists in living force, so that it is neither subjective nor objective, but the unity of them both. Third, through its organic structure all of nature conforms to a purpose, plan, or design, which is not created by God but inherent in matter itself. The first proposition makes absolute idealism a form of monism; the second makes it a form of vitalism; and the third makes it a species of idealism" (Beiser, 2000, 34). In a general way, one may, following these considerations, characterize Hegel's philosophical system as a form of Absolute Idealism, the single substance being *spirit*, the "living force" being *will*, and the "purpose, plan or design" inherent in matter being *self-cognition*.

8. We may speak, without contrition of "consciousness" too, but we prefer to avoid such Philosophical territory, in order to put forward our main argument, concerning the Hegelian notion of *free will*. Nevertheless, Hegel's definition of consciousness is as follows: "Consciousness (*Bewusstsein*), in general, is the relation of the I (*Ich*) to an object (*Gegenstand*), whether it is an internal or an external object. " (Hegel, 1986, 204). This early Hegelian definition of consciousness is surprisingly similar to that explored by Husserl and Husserlian Phenomenology in the XXth century: "One could say that wherever and however there exists that which I call consciousness (*conciencia*), I find it always constituted by two elements: an attitude or act of a subject, and a 'something' toward which such act is directed to" (Ortega y Gasset, 1963, 62).

228).[9] So, a natural being or thing, is, strictly speaking, that which cannot be made something other from what it already is, except by some *external* intervention.[10]

Concerning spirit, "the most sublime concept of all" (Hegel, 1986b, 28), the following passage from the 1824/25 lecture on the Philosophy of Right is helpful: "Spirit (*Geist*) is thinking (*Denken*) in general, and man is different from the animal through thought" (Hegel, 1974b, 102). Though Hegel does not offer here a wide definition of "spirit" (which, in any case, is the philosophical task of the *Phenomenology of Spirit* and the whole Hegelian system), at least we get a sense, of where the nucleus of the question lies: spirit, contrary to nature, is not a *given*, an immediate "reality", rather, it is a process, and a result as well (this is an essential point in the *Phenomenology*[11]); "of what?" may the reader ask; to which the Hegelian answer states: *of thought*. This standpoint leads to understandings and theses such as: *spirit is its own concept presented in and through thought*, *spirit is self-thinking thought*, *thought as subject and object as well, is spirit;*[12] now, we must decisively state that this sentences do not constitute mere tautologies, pseudo-propositions or mad jabber, such as Schopenhauer, Carnap and others would urge us to conclude;[13] on the contrary, they are nothing but a succinct and *general* summary of the whole architectonic of concepts and argumentations in which the Hegelian system consists of. It would be only just to conclude that *spirit* is a set of logical, physical,

9. Along this paper we shall make use of several Hegelian sources on the Philosophy of Right; that is, apart from the "print" *Philosophy of Right*, we shall make use of manuscripts of the Hegelian courses on Philosophy of Right, which the German philosopher imparted between 1817 and 1831, and which were, though in a fragmentary and sometimes incomplete way, "recorded" by students.

10. One can now understand and even defend the Hegelian argument which states, that it is precisely the role of culture (*Bildung*) to introduce from *outside into* the children the determinations of the spiritual. A child without culture would be pretty much just an animal, such as is, in the beginning, the feral lad of Werner Herzog's *The Enigma of Caspar Hauser*. Interestingly Hegel and his family were well acquainted with the Kaspar Hauser *affaire* around 1829; Hegel's mother-in-law, Susanne von Tucher, wrote in that year to the philosopher's wife: "Kaspar thanks you for your interest, of which I have told him" (Tucher *apud* Beyer, 1966, 101).

11. "The True is the Whole. The Whole, however, is only the essence which realizes itself through its development. From the Absolute it must be said that it is essentially a *result*, that it only in the *end* (*Ende*) is, that which it truly is" (Hegel, 1986b, 24).

12. The reader may analyze the whole Hegelian *corpus* and indeed find these very same sentences stated as such, in one way or the other.

13. See Schopenhauer's *On the Philosophy at the University* and Carnap's *The overcoming of Metaphysics through the logical analysis of language*.

chemical, biological, anthropological, psychological, juridical, moral, economical, political, aesthetical, religious, philosophical, and historical categories, which *explain* what *reality* is (which at the same time cannot but explain what *man* is). In Hegel words: "Spirit is not an abstract thing; it is, essentially, a system which differentiates within itself" (Hegel, 1983b, 64). A system *of concepts*, which, one may consider, is presented in an utmost summarily way in the Hegelian *Encyclopedia*.[14]

Now, we may return to that previous position which states than man "*is not from nature, what he should be;*" indeed, if there is an *essential* difference, between plants, animal and man, is this deontological character *unique* to man. In this sense, man not only *is*, but possesses (or conceives, which would be pretty much the same, in the Hegelian view), a certain *ideal*, archetypical, or deontological dimension, which constrains his own actions, as well as places him within a determinate frame of *valid* interaction with other human beings; in Hegel's own words: "The animal is not in conflict with what it should be; man, on the other side, must know precisely that: what he should do, to conceive what he should do; and, in this way, to give his will (*Willen*), his sentiments (*Gefühlen*), his impulses (*Trieben*), a true content" (Hegel, 1974a, 495). To state it simply: human beings, *necessarily*, conceive a self-definition as agents, and a frame of acceptable social behavior: naturally, there is a historical side of this self-definition and social behavior, which is explained by Hegel in terms of *evolution of the human mind*, or which is the same, the "development of spirit" (Hegel, 1991, 315).

Now, this leads us to a decisive step in our argument, namely, the definition of "will" and of such a thing as a "true content" for this will. Notoriously there is an utmost complex philosophical explication (at least more complex than that which one may find in Hobbes, Spinoza and Kant, three great figures in the Modern debate concerning *freedom of the will*) concerning the matter, which is developed in the *Philosophy of Right* in paragraphs §5, §6 and §7, following an argumentative development closely based upon theoretical background from the *Science of Logic*, as Klau Vieweg defends;[15] in Hegel's words: "We must now consider: 1. will in general, 2. particular will, natural and

14. Concomitantly one may consider that the Hegelian lectures on the Philosophy of Right, the Philosophy of Art, the Philosophy of Religion, the History of Philosophy, and the Philosophy of History are wider presentations of the very same subject matter of the *Enciclopedia*; and as such, the ultimate subject matter in them is *spirit*.

15. "The fundamental determinatedness of the concept of the free will as 'principle and beginning of the Science of Right' can only be inferred and understood in connection to Hegel's innovative logic" (Vieweg, 2012, 57).

reflexive will, 3. free will in and for itself, which determines itself also; nonetheless in its determinacy this will remains a really free will. §5, §6 and §7 give this moments" (Hegel, 2012, 43). So, Hegel means that such thing as a will[16] has a tripartite character: first, the possibility of abstracting from any specific content: second, the possibility of resolution to action (that means the abandonment of the tenacity of abstraction and refusal to action); third, the possibility of taking as an action guideline or principle, not just the satisfaction of impulses or desires, being what they may (Hegel would speak of *finite ends*[17]), but that which expresses the very essence or character of will, namely, concrete universality (Hegel would speak of *infinite ends*[18]). The reader may indeed miss the clarity and precision of Hobbes, Spinoza and Kant, but it is also indeed *impossible* at all, to extract from Hegel a "simple" definition, which would not *necessarily* end up in a plea for *universality* or for articulation into a wider *system of concepts*. So, for the sake of understanding the gist of Hegel's theory of the will, let us add some additional remarks, in order to give a final conclusion to the matter.

"The will, as the interior determinant concept, is essentially *activity* (*Tätigkeit*) and *action* (*Handlung*). It translates its interior determinations into an exterior existence, in order to present itself as Idea"[19] (Hegel, 1986a, 57). Apart from the definition of will that we obtain here ("the interior determinant concept") we get an intensive view on the nucleus of the Hegelian theory of the will, namely, the thesis of the *translation* of "interior determinations into an exterior existence", in order words, that the essence of the will is *action*,[20] understood as a change in the world caused, not just by chance, or by blind movement, but by a certain intentionality and causation-by-agent; "To *action* (*Handlung*) belongs, above all, only that, which was in the decision (*Entschluss*) or consciousness" (*ibid*.) states Hegel, intending precisely to argue that in action, properly

16. That means: *human will*. Hegel would concede the use of the *coniunctum verborum* "animal will", if by such it is understood simply: instinct and avidity.

17. See Hegel (1979, 277)

18. See Hegel (1983a, 110)

19. This fragment comes from a philosophical Encyclopedia for Gymnasium students. Together, with the philosophical Handbook on rights and duties quoted above, this is the most didactic and synthetic exposition of the Hegelian thought concerning ethics and political philosophy. *The Philosophy of Right* may be more systematic, developed and exhaustive, but it lacks the freshness and pedagogical character of this two earlier ethical and political sources.

20. Hegel states in the Hotho manuscript from the course on Philosophy of Right of 1822/23: "A will, which does not decide or resolves itself to action, is not a real will (*wirklicher Wille*)" (Hegel, 1974a, 130).

speaking, it is decisive the intending, of the acting agent, of causation of a change in the external world, in other words, an action is a product of a decision or resolution of an acting agent, and this resolution takes place *in* him, so one must unconditionally conclude, that such a resolution *belongs* to him, that such a resolution is *his* and no one else's. One may also guess, that this theoretical position has relevant juridical and moral consequences (indeed, such is the very same Hegelian understanding of the matter, as we will see later).

We may now state a concise Hegelian summary of the arguments presented until this point: "All determinations of the will can be called ends or purposes (*Zwecke*), determinations, which should be valid" (Hegel, 1979, 59). This statement is decisive, and as a matter of fact, marks the acute controversy which Hegel maintains with Kantian Idealism[21]: will has *necessarily* determinations or specifications (that is precisely the logical-philosophical argument on paragraph §6 of the *Philosophy of Right* – will cannot stay at an abstract point of indeterminacy, it must, *resolve* to *something*), and inherent to any such specification is *action*, that is, to make the external meet the requirements of the internal, the world to the ends.[22] Hegel speaks of "natural will (*natürliche Wille*)", "arbitrariness (*Willkür*)", and "reflexive will (*reflektierende Wille*)", (Hegel, 1983a, 217) precisely to describe this mind-to-world direction of human agency, to express it in contemporary terms; the decisive here is that what is as stake is the satisfaction, of impulses or desires ("natural", or "artificial"[23]), through determinate means, the

21. "The particular will should be adequate to the universal Will, this unity is postulated; man should be moral, but this stays at the level of a mere *should* (*Sollen*) […] We stay here, therefore, on the level of a mere talk about morality" (Hegel, 1986c, , 369). So, Hegel's ultimate controversy with Kantian Ethics lies in this simple question: is it enough that a subject or a will *intends* to do "the good" without doing *anything at all*, in a concrete way? Hegel rejects totally the *practical* relevance of such moral individual deliberations. To state it in a Goethian *dictum* with which Hegel completely agrees: *man is what he does*.

22. Hegel expresses this in this way: "Through *acting*, the interior practical determinations (*innerlichen praktischen Bestimmungen*) obtain an exteriority, that is, an exterior existence. Inversely, this can be considered in this way: an immediate exteriority is cancelled, and is made concordant to the interior determination" (Hegel, 1986a, 205).

23. Notoriously, Hegel refuses to draw a clear and distinct line between the natural and the artificial *qua* desires, impulses or needs (*Bedürfnisse*); this standpoint has relevant economic and philosophical consequences, as it implies that the multiplication of needs and means to satisfy them, peculiar to the Modern World, far from being a matter of lamentation and diatribes *à la* Rousseau, is a matter of economic and philosophical celebration, as it means the conquest of spirit (or man, which is the same) over nature, in the material and the intellectual. (See §190 of the *Philosophy of Right* and its equivalents in the manuscripts of Philosophy of Right).

specifically human in the matter consisting in the *reflective* character of the action, which in simple terms signifies the choosing of the adequate *means* to the *ends*, and being in the possibility of eventually, *refraining* from action[24]. On a simple (finite, in Hegelian terms) level that is what makes human *actions* different from animal *instinctive* engagements with the world.

On a complex (infinite, in Hegelian terms) level, it is required, not only that a will resolves itself to action or that it renounces execution of ends; indeed, the end of an agent may be this or that, to buy chocolate ice cream or an opera ticket, to accept a job or to rob a bank. At this level, there would not be *any* criterion *at all*, to discriminate between ends, means, or actions *qua* validity, acceptability, legality, etc. The argumentative step sketched above as "free will in and for itself", and "really free will", the subject matter or paragraph §7 of the *Philosophy of Right*, intends precisely to argue for the necessity of such a task. Paragraph §33 of the same text defines concretely what must be understood as the determinations of "really free will:" "According to the process-steps of the development of the Idea of the free will in and for itself, is the will […] *formal right* (*formellen Rechts*) […] *morality* (*Moralität*) [and] *ethical life* (*Sittlichkeit*) [which in itself develops into] *family* […] *civil society* (*bürgerliche Gesellschaft*) [and] *State*" (Hegel, 1979, 88). In simple terms, aside from the fact that human beings conceive ends and execute them or reject them, there is, and *must be*, a sphere of *normativity* which establishes what a valid/invalid action (*qua* execution of ends) is. That is the main task of the *Philosophy of Right*, to define categories such as property, the morally good, family-care, economic production and division of political powers, which, in the Hegelian perspective, are, in the end, categories of the *free will*.

With this last elements, we could now "easily" understand such terms as "*inferior desire faculty* (*niedere Begehrungsvermögen*)" (Hegel, 1986a, 205) and "*superior desire faculty* (*höhere Begehrungsvermögen*)" (Hegel, 1986a, 206); the first points out to the satisfaction of *any* end, independently of any determinate normativity, and the second to the fulfillment of such ends, which are concordant to the essence of character of the will itself. This last point establishes the most attractive and original, as well as polemic, argument of Hegel concerning the free will; in his own words: "In the usual

24. Paragraph §5 of the *Pflichtenlehre* states it this way: "Through *reflection* (*Reflexion*) man goes beyond impulse (*Trieb*) und its limitations. He compares this impulse not only with the means of its satisfaction, but also this means with one another, and impulses with one another; and also with the ends of his being, and allows himself the conclusion of reflection, whether as a satisfaction of the impulse, or as its detention, and renunciation" (Hegel, 1986a, 206).

representation, will and intelligence appear as two distinct things. Free will, however, which has as content nothing else as itself, has its content only through thought" (Hegel, 1983b, 64). So, Hegel's ultimate argument for the definition or, existence even, of "really free will" lies in this thoroughly metaphysical (in the sense of non-empirical) standpoint: when the will establishes as its content, not just any thing, not just any end, but itself, and the end of understanding and explaining itself as free, as non-natural, as non-animal, we have *in actu* the existence *par excellence* of such a thing as *freedom*; that is the background for the understanding of the central argument of the Hegelian Philosophy of History: "The oriental peoples [that is peoples with a patriarchal and despotic social structure] do not know that spirit, or man as such, is in itself free. As they do not know it, they are not free" (Hegel, 1986d, 31). To be free, strictly speaking, one must *know* himself to be free, and concomitantly, to will and argue for the *known* (*erkannt*) concrete determinations, and institutions of free will: private property, economic freedom,[25] and State founded on the rule of law (*Rechtsstaat*) are to be, according to this view, wished for and defended in order to give *free will* actuality and concrete validity.

In conclusion, the Hegelian theory of free will, does not consist of a discussion (such as is the case in Hobbes, Spinoza and Kant) of the validity or invalidity of such a concept as "uncaused cause", "effect without previous cause" or "independence from causation;" on the contrary Hegel plainly states that will is indeed "determined by something", and as such is "unfree;"[26] Hegel's contribution to the Old Question[27] of the freedom of the will lies precisely in the redefinition of the *coiniunctum verborum* "free will:" a free

25. The relationship of Hegel to Political Economy and Capitalism, in general, is complex indeed. Let us summarily state that Hegel accepts the economic category of *capital*, with the liberty of production and consumption, which it implies; nevertheless he argues for social and state institutions which safeguard human dignity as a moral and ethical agent.

26. "The will is determined (*bestimmt*) by something; therefore the will is not free (*ist nicht frei*)" (Hegel, 2012, 50).

27. By "old" here, we mean simply, present, in one way or the other, since Ancient times, for instance in the now famous thesis from Lucretius of the "swerving atoms" non subject to mechanical causality: "Again, if all movement is always interconnected, the new arising from the old in a determinate order – if the atoms never swerve so as to originate some new movement that will snap the bonds of fate, the everlasting sequence of cause and effect – what is the source of the free will possessed by living things throughout the earth?" (Lucretius *apud* Dennet, 1984, 2). On the other side, strictly speaking, the concise stating of the problem of the free will, concerning a non-theological Ontology (which would be the ultimate standpoint in Saint Augustine and Aquinas), is reached only in the *Modern World*, in *Modern Philosophy*. So, the conceptual and systematic tackling of the problem of the freedom of the will is only reached from Descartes onwards.

will, is a will which has ends, and which understands or cognizes itself, as being such an ontological instance which defines *from itself* what it iself *is* and *should be*. That this result should be arrived at *necessarily*, that is, without the intervention of a noumenical Kantian independent-of-causality faculty of ends, is, in the final picture, irrelevant for our German philosopher.

If there ever was a pure compatibilist[28] philosophical scheme of Philosophy of Mind, concerning freedom and causality, it is the Hegelian theory of the free will.

II

> In the present I am conscious of my reality (*Wirklichkeit*); and consequently self-consciousness find itself as matter – the soul as material, mental representations as movements and changes in the interior organ of the brain, which follow after impressions of the senses. (Hegel, 1986c, 289)

This fragment from the Hegelian lectures on the History of Philosophy seems, at first glance, to defend a strict materialist Ontology, which would be not incompatible with contemporary discussions on Philosophy on Mind;[29] unfortunately, we must, declare that this Hegelian argumentation occurs in a theoretical *locus* which intends, *precisely*, to denounce and criticize the partiality and unilaterality of materialist and atheistic Enlightenment Philosophies (such as Holbach's and La Mettrie's), in order to defend the centrality of categories such as *cognition*, *free will*, *right*, etc., in reality in general. Nevertheless, this allusion of Hegel to the "interior organ of the brain" is not unimportant; it shows Hegel's genuine interest in Physiology, which is also evidenced in his *Encyclopedia* explications on the "animal organism" (Hegel, 1991, 291) and the "system of embodiment of the spiritual" in man (Hegel, 1991, 328), or, in other words, with actual *organs* and *physiological systems* which underlie behavioral and cognitive processes. So, it is not that, in the end, Hegel considers Anatomy and Physiology to be

28. Compatibilism taken to mean the philosophical acceptance of the notion of "physical" or "mental" *phaeonomena* as being absolutely subjected to causation and the acceptance, as well, of the validity of the category of *freedom of the will*. In the terms of a student's manual on Philosophy of Mind: "Compatibilism says that the up-to-us-ness of our actions – our freedom to act otherwise – is entirely compatible with our actions having been all along predetermined by causes outside our control. Freedom and causal determinism are perfectly consistent" (Pink, 2004, 19).

29. For example, Daniel Wegner, Benjamin Libet, Daniel Dennet and John Searle.

irrelevant to philosophical inquiry; rather, Hegel's decisive thesis, in this matter, is simply: the empirical study of nature (which would include, in a Hegelian-inspired Ontology such entities as the corpus callosum and the limbic system), in itself important, cannot yield a complete understanding of reality, as in reality, there are elements, which are not merely *given*, but are *produced* by man's engagement with the world. Indeed, one may have a global account of the physical and chemical constitution of the world, and yet find not a single glimpse of a *right*, *the moral good*, *economic capital* and *political sovereignty*; all this instances are not material but *spiritual*, which simply means, again, that they are a result of human activity, and not merely of the execution of the DNA program inside the human cells.

In this sense, the Hegelian theory of the free will (and his whole philosophical system as well) is of great relevance to contemporary discussions on Philosophy of Mind, Philosophy of Right, Bioethics and Neuroscience.

Hegelian arguments concerning free will can be, to some extent, experimented and observed in the laboratory, as his monistic Ontology does not exclude the possibility that the mental can be instantiated in the material (on the contrary, Hegelian dialectics can be interpreted as a doctrine of the mutual interaction of the material and the mental). Spinoza argued for the *strict qualitative difference* between the material and the mental; Kant argued for the noumenical (that means, in one sense, *suprasensible*) character of the will, which may be compatible with such a thing as an *uncaused initiator of causal series* in the empirical world; in both schemes of thought neurological research on the subject of the free will is irrelevant. In Spinoza because *de facto* we already know that every material event has a previous cause (so necessarily we must exclude the existence of such a thing as a material or mental effect without a previous cause), in Kant because neurological experimental techniques can only yield information in the frame of the empirical, and as such, in the frame of phaenomena constituted *a priori* by our own mental faculties, the noumenical world remaining *absolutely* closed to experimental research. Hobbes may effectively yield a thought frame compatible with neurological categories, and his definition of will as "*last Appetite in deliberating*" (Hobbes, 1929, 47) can be tested experimentally as prefrontal cortex activity preceding motor activity; concerning the question of free will, Wegner's and Libet's experiments would be completely compatible with the Hobbesian rejection of the notion of a will safeguarded from exterior or empirical influence (that is, free will understood as "uncaused cause").

On the other side, Hegel's "inferior faculty of desire" can be specially be subject for experiment and observation; for instance, that the prefrontal cortex is responsible of such "human intellectual traits" as "*judgement*, *foresight*, a sense of *purpose*, a sense of

responsibility and a sense of *social propriety*" (Haines, 2002, 518) is a widely accepted fact in contemporary Neurobiology. This is a concrete point in which Hegelian Philosophy and Neurobiology could come together to render some applicable results on Neuroethics and Philosophy of Right; indeed, if free will, and concomitantly teleological (even if on a finite level) behavior is inherent and essential to man, and if this trait is the philosophical basis for right and jurisprudence, then, a clinical case, with impaired teleological activity due to some physiological abnormality (produced by genetics, tumors, infections, etc.) in the structures[30] which are (to our present knowledge) conditions *sine qua non* for "executive function"[31] should, in a strong normative sense, be susceptible to particular juridical treatment, as that which make man *human*, that is *free will*, would be missing, in some way or extent; to state it briefly, a human being with limited teleological capacity should be, concomitantly, a subject with limited juridical capacity. The special juridical treatment of children and the mentally ill or impaired is philosophically justified in a strong and empirically verifiable way.

Concerning the Hegelian "superior faculty of desire" the task may not be so easy as in the "my brain, my action, my responsibility" case stated above; let us just, in a challenging brief way, state that a "left lateral prefrontal gioblastoma" case described by Knight and D'Esposito resulting in impaired social capacity,[32] could, in a global sense, just be judged in view of the "really free will" invoked by Hegel on the question of the *superior* use of our teleological capacity; indeed, the thesis that being part of a social community, in a functional as well as "healthy" way, is desirable in and of itself, is something that could scarcely be accepted philosophically without an ontological background such as the one in the theory of the free will of the author of the *Philosophy of Right*. Again, one may have a complete account of the functioning of the whole brain circuitry at the genetic, anatomical and electrochemical level, and yet find no glimpse of *rights*, *the moral good*, *economic capital*, and *political sovereignty*.

Finally, the question for the freedom of the will, which occupied modern philosophers such as Descartes, Spinoza, Hobbes and Kant in an intensive way, received in Hegel a

30. Drubach *et al*. speak of "the dorsolateral prefrontal cortex (DLPF) and the ACC [anterior cingulated cortex] " as brain areas "most frequently implicated in control of executive behavior" (Drubach *et al*, 2011, 245).

31. Knight and D' Esposito define "executive function" as "a wide range of cognitive processes such as focused and sustained attention, fluency and flexibility of thought in the generation of solutions to novel problems, and planning and regulating adaptive and goal-directed behavior" (Knight and D' Esposito, 2003, 259).

32. "[W.R.] was unable to carry out the activities necessary to make him a fully functioning member of society" (Knight and D' Esporito, 2003, 261).

radical reorientation: it is not after an "uncaused cause" which we should look or yearn for, in order to feel and think ourselves as *free*; rather the very same material-biological reality, and, as such, thoroughly submitted to causality, leads us *necessarily* to conclude that we, as human beings, are not given entities submitted to the fate of the execution of a genetic or algorithmic program; on the contrary, we make ourselves, as individuals and collectives, something *other* than that which our DNA dictates (that is: survival, in a general sense). To this self-production, self-assertion, eventual self-recognition of man through man, and definition of the *essential* in him Hegel calls *free will*.

Bibliography

Beiser, Frederick. 2000. "The Enlightenment and idealism." In *The Cambridge Companion to German Idealism*, edited by Karl Ameriks. Cambridge: Cambridge University Press.

Beyer, Wilhelm R. 1966. "Aus Hegels Familienleben." In *Hegel-Jahrbuch*, edited by Wilhelm R. Beyer. Germany: Anton Hain.

Dennet, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford: Oxford University Press.

Drubach, Daniel, Alejandro Rabinstein, and Jennifer Molano. 2011. "Free will, freedom of choice and Frontotemporal Lobar Degeneration." *MSM* 9 (1): 238-250.

Haines, Duane E. 2002. *Fundamental Neuroscience*. 2nd ed. London: Churchill Livingstone.

Hegel, G.W.F. 1983a. *Die Philosophie des Rechts*. Die Mitschriften Wannenmann (Heidelberg 1817/18) und Homeyer (Berlin 1818/1819. Stuttgart: Klett-Cotta.

Hegel, G.W.F. 1983b. *Philosophie des Rechts, Die Vorlesung von 1819/20 in einer Nachschrift*. Germany: Suhrkamp.

Hegel, G.W.F. 1974a. *Vorlesungen über Rechtsphilosophie (1818-1831)*. Vol. 3. Stuttgart: Frommann-holzboog.

Hegel, G.W.F. 1974b. *Vorlesungen über Rechtsphilosophie (1818-1831)*. Vol. 4. Stuttgart: Frommann-holzboog.

Hegel, G.W.F. 1979. *Grundlinien der Philosophie des Rechts*. Germany: Suhrkamp.

Hegel, G.W.F. 1986a. *Nürnberger und Heidelberger Schriften 1808-1817*. Germany: Shurkamp.

Hegel, G.W.F. 1986b. *Phänomenologie des Geistes*. Germany: Suhrkamp.

Hegel, G.W.F. 1986c. *Vorlesungen über die Geschichte der Philosophie III*. Germany: Suhrkamp.

Hegel, G.W.F. 1986d. *Vorlesungen über die Philosophie der Geschichte*. Germany: Suhrkamp.

Hegel, G.W.F. 1991. *Enzyklopädie der philosophischen Wissenschaften (1830)*. Germany: Felix Meiner Verlag.

Hegel, G.W.F. 2012. *Die Philosophie des Rechts, Vorlesung von 1821/22*. 2nd ed. Germany: Suhrkamp.

Hobbes, Thomas. 1929. *Leviathan*. Oxford: The Clarendon Press.

Knight, Robert T., and Mark D'Esposito. 2003. "Lateral Prefrontal Syndrome: A Disorder of Executive Control." In *Neurological Foundations of Cognitive Science*, edited by Mark D'Esposito. Cambridge: Massachusetts Institute of Technology.

Ortega y Gasset, José. 1963. *Obras Completas*. Tomo II. 6th ed. Madrid: Revista de Occidente.

Pink, Thomas. 2004. *Free Will: A very short introduction*. Oxford: Oxford University Press.

Vieweg, Klaus. 2012. *Das Denken der Freiheit: Hegels Gundlinien der Philosophie des Rechts*. München: Wilhelm Fink Verlag.

Journal of Cognition and Neuroethics

Collecting Evidence for the Permanent Coexistence of Parallel Realities: An Interdisciplinary Approach

Christian D. Schade
Humboldt-Universität zu Berlin

Biography

Christian D. Schade is a full professor at Humboldt University's School of Business and Economics and Director of the Institute for Entrepreneurial Studies and Innovation Management. Furthermore, he is a Research Fellow at Wharton's Risk Management and Decision Processes Center (University of Pennsylvania). His research contributes to a better understanding of decision making in general and of entrepreneurial as well as innovative decision making. He is currently working on novel foundations and perspectives for the decision sciences. His research is mainly based on laboratory experiments, economic psychology and mathematical psychology, as well as quantum mechanics.

Acknowledgements

I am thankful to Michael Mensky for the inspiring discussions we had on his many-worlds interpretation of quantum mechanics. I am grateful to Tanja Strohm for the intellectually challenging conversations we had on practically all topics of the paper as well as for her many helpful suggestions and comments on the manuscript. I thank Anna Abratis for her contribution to the literature search and her as well as Christine Lauritzen and Cristian Stefan for their comments on an early version of the manuscript. I am grateful to Elisabeth Karsten for her suggestions pertaining stylistic changes as well as extensions of my thoughts. I furthermore thank the participants in a long presentation at Ludwig-Maximilians-Universität in Munich, especially Markus Maier, for their important questions and comments. I finally thank the participants in my presentation at IINN's Free Will Conference in Flint, Michigan, especially Shai Frogel, Robert Oszust, and Adam Taylor, for their great questions and suggestions.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Schade, Christian D. 2015. "Collecting Evidence for the Permanent Coexistence of Parallel Realities: An Interdisciplinary Approach." *Journal of Cognition and Neuroethics* 3 (1): 327–362.

Collecting Evidence for the Permanent Coexistence of Parallel Realities: An Interdisciplinary Approach

Christian D. Schade

Abstract

This paper assembles an interdisciplinary 'presumptive evidence proof' for the existence of parallel worlds, hence supports physics in solving the interpretation problem of quantum mechanics by making use of theory and experimental findings from psychology, philosophy, and the neurosciences. It will demonstrate that two questions are closely intertwined: the question of whether an *actual free will* exists and the *interpretation of quantum mechanics* chosen. Specifically, the paper will argue that whereas there is no room for an actual free will if the *Copenhagen interpretation* of quantum mechanics (postulating the 'disappearance' of Schrödinger's wave function and the appearance of a singular state under measurement) is adopted (the same is true with other interpretations involving a collapse of the wave packet), an actual free will is possible if specific versions of the *multiverse interpretation* are chosen. This point cannot be made directly. In fact, it can *only* be produced within the proposed, interdisciplinary 'presumptive evidence proof' for the coexistence of parallel realities. Finally, the paper tentatively suggests an 'interpretation' of the many-worlds interpretation that circumvents some of the 'strange' ontological implications that this perspective exhibits according to some of its previous interpretations and develops a view on how free choices might actually be made.

Keywords

Parallel realities, free will, consciousness, many-worlds interpretation of quantum mechanics, Copenhagen interpretation of quantum mechanics, time, decision making

Interpretation of quantum mechanics as an interdisciplinary effort

Numerous interpretations of quantum mechanics have been proposed, and the theoretically most coherent – but also most thought provoking – of them, the many-worlds interpretation by Everett-DeWitt (Everett 1957; DeWitt 1970, 1971), or any other, more recent version of the multiverse view, would have huge consequences for our worldview also outside physics. Building conclusive evidence for any of the multiverse interpretations within theoretical and experimental physics alone is rather hard or perhaps even impossible at this point. Therefore it is important to take into account theory and experiments also from other scientific domains that are of fundamental relevance in this regard.

Consequently, the paper is assembling a 'presumptive evidence proof' to go as far as possible with making the permanent coexistence of parallel realities plausible. Specifically, the paper crafts an interdisciplinary approach, predominately based on physics, psychology, neuroscience, and philosophy. In the center of the argument are considerations on *free will*. According to any version of the multiverse view, *different realities* permanently coexist. Whereas this view uncomfortably suggests that our everyday experiences are based on a somewhat limited (or at least incomplete) picture of the actual world, other interpretations of quantum mechanics also come at a price.

As the paper is going to argue, other interpretations of quantum mechanics such as the popular Copenhagen interpretation – postulating a 'wave function collapse' resulting in a singular reality – are inconsistent with the existence of an *actual* free will; whose proposed absence is intuitively rejected by most people outside science (Nichols 2011). Interestingly then, the line of arguments Hameroff (2012) quite recently presented in favor of an existence of free will in light of quantum brain biology will turn out to be partially related to the respective argument presented in this paper on the one hand. However, on the other hand, whereas Hameroff (2012) argues that the objective reduction modification of quantum mechanics (Penrose 1994; Hameroff and Penrose 1995) – a singular-universe approach – would also be able to 'rescue' free will, this paper will argue that the latter is only possible in the multiverse. More precisely, this paper will show that quantum mechanics *is* free will friendly, but only if we (a) allow for the permanent coexistence of parallel realities and (b) if specific 'interpretations' of the many-worlds interpretation are chosen. As our analysis is going to demonstrate, one of the existing multiverse interpretations (the EEC by Mensky 2005, 2007a,b, 2010) is indeed free will friendly. It will turn out, however, that this approach has strange consequences, ontologically, as any other of the existing multiverse interpretations to be analyzed in this contribution. Hence the paper will tentatively propose a new interpretation of the multiverse whose consequences might be seen as ontologically less irritating. The paper will finally address the question how free choices might be made, what it actually means to freely choose between alternatives in the multiverse.

The contribution is building up primarily on the seminal works by David Deutsch (Department of Atomic and Laser Physics, Centre for Quantum Computation, Clarendon Laboratory, Oxford) and Michael Mensky (Lebedev Physical Institute, Russian Academy of Sciences, Moscow), both very outspoken about their preference for a multiverse interpretation of quantum mechanics, and both publishing their thoughts in scholarly journals as well as popular science monographs (for the latter see, e.g., Deutsch 1997; Mensky 2010).

Regarding the contributions by Deutsch, the paper is sharing many of his thoughts on the nature of time (see the proof section of the paper, step 3). Other basic premises of the contribution are related to the work by Mensky, he himself mainly building up on the work by Squires (1988). According to Mensky (2010, 54), essential arguments against von Neumann's ([1932] 1996) reduction postulate, explicating the Copenhagen interpretation, "will be connected with the phenomenon of consciousness." Hence, the idea that only an *interdisciplinary treatment* may suffice in generating a convincing case for the many-worlds view can be traced back to the works by Mensky (e.g., 2005, 2007a,b, 2010).[1]

The 'presumptive evidence proof' for many worlds is presented in a stepwise manner, as *pieces of a puzzle* that will finally form a coherent picture. The pieces of the puzzle are taken from different domains, mainly quantum mechanics (measurement/interpretation problem etc.), role of consciousness (in quantum measurement as well as in light of neuroscience findings), findings on/explanations of predictive physiological anticipation, and considerations on the possibility of free will (being at the core of the contribution). Interestingly, *within each* of these *domains* (i.e., quantum mechanics, free-will problem in philosophy, etc.) there are *alternatives* to treat or interpret the respective phenomena or theories, but the flexibility is gone when trying to form a joint perspective out of all those domains. Indeed, within each of those disciplines there is always just one approach that qualifies as piece of a puzzle appropriate to complete the picture. This idea is depicted in Figure 1.

The remainder of the contribution is structured as follows. In the next, main chapter, the paper will craft, in a stepwise manner, a 'presumptive evidence proof' for parallel existing realities. It ends with a long subchapter on free will and on interpreting the many-worlds interpretation in a form that is free will friendly and makes sense, ontologically. This chapter is followed by a chapter addressing the question how free choices are made in the multiverse. The final chapter contains a conclusion and remarks,

1. However, whereas Mensky's work is firmly rooted in the measurement theory of quantum mechanics, his psychological arguments are rather presented in the form of 'anecdotal evidence.' Instead, the goal of this paper is to push as much as possible towards a 'proof,' given the interdisciplinary knowledge we have. This requires being as specific, as rigorous with arguments from psychology, philosophy, and the social sciences as with those from quantum mechanics. Furthermore, this paper is going to reverse some of Mensky's arguments. What he sometimes postulates for the sphere outside physics, this contribution shall employ together with (additional) empirical or theoretical evidence in favor of the respective phenomena, to substantiate the parallel and permanent coexistence of multiple parallel realities.

it summarizes the results of the presented analysis, and speculates on what consequences the adoption of a multiple-realities perspective might have.

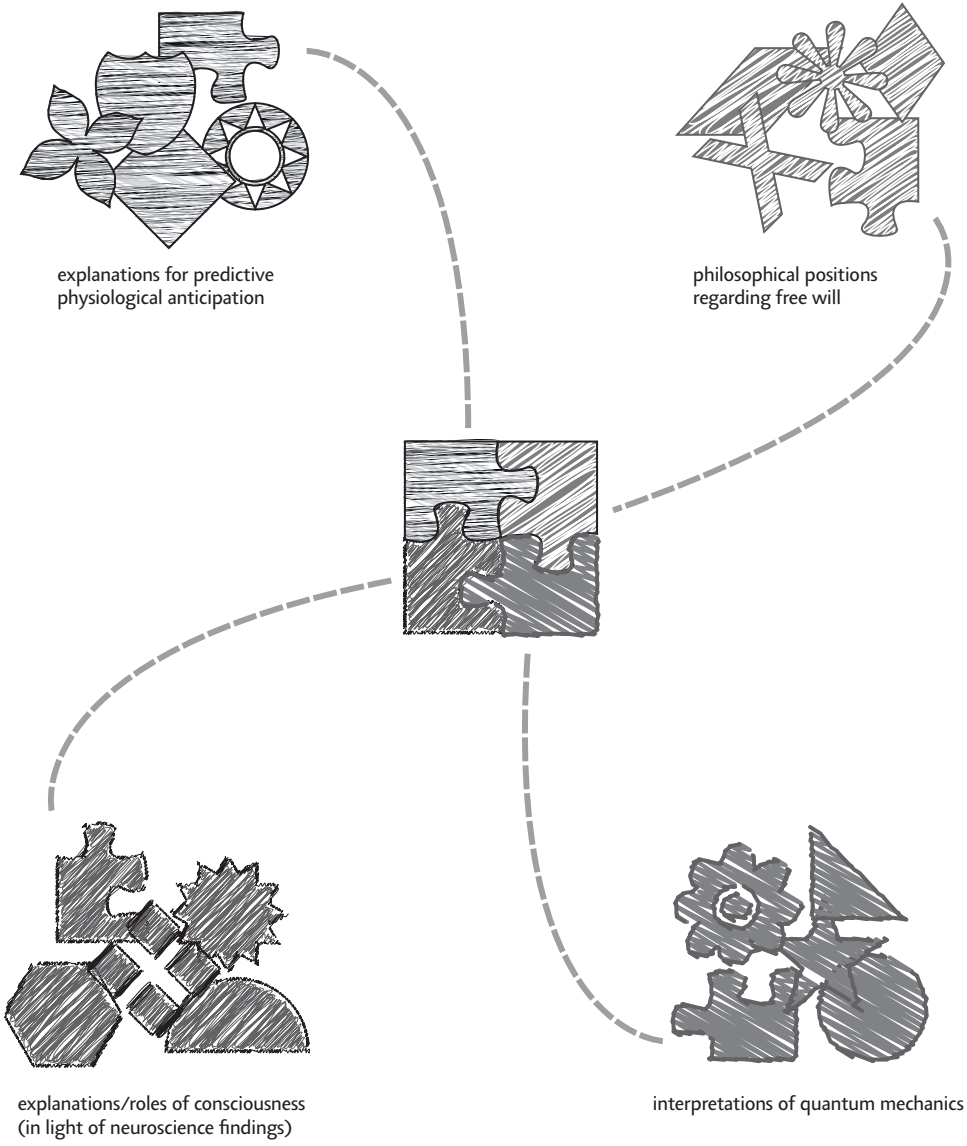


Figure 1: Structure of the interdisciplinary 'proof' of the multiverse

A 'presumptive evidence proof' for the coexistence of parallel worlds

Step 1: Many worlds as a convincing interpretation of quantum mechanics

Of the many possible interpretations of quantum mechanics (see, e.g., Auletta 2001), all being trials to address the so-called measurement problem,[2] the most well-accepted so far have been the Copenhagen interpretation[3] (together with von Neumann's reduction postulate (von Neumann [1932] 1996)[4]) as well as the many-worlds interpretation, initially based on Hugh Everett (1957) and its further interpretation by Bruce DeWitt (1970, 1971).[5] In this article, I shall mainly concentrate on those two; although the objective reduction formalism[6] (Penrose 1994; Hameroff and Penrose 1995) will briefly be touched, and, as already mentioned, the many-worlds interpretation will again turn out to be interpretable.[7] Dealing with other interpretations is beyond the scope of this article.[8]

2. An important aspect of the measurement problem is the fact that measurement results achieved on some quantum system are uninterpretable without taking into account the consciousness of the observer. This turns out to always be the end of a logical chain of reasoning defining a measuring device, then defining the brain as evaluating the result shown on the measurement device, etc.

3. The Copenhagen interpretation of quantum mechanics is the one most representative of something that might be called a 'quasi-Newtonian' worldview; it is that interpretation of quantum mechanics that challenges the validity of our everyday experience in the least radical way.

4. For simplicity, whenever the paper mentions the Copenhagen interpretation, this (most prominent) version of it is meant.

5. Within the academic community in physics, other well-known current or past proponents of the many-worlds interpretation are, e.g., David Deutsch, Murray Gell-Mann, Richard Feynman, Stephen Hawking, Michael Mensky, and Euan Squires.

6. Since objective reduction changes the formalism of quantum mechanics, it is actually more than an interpretation.

7. Examples for interpretations of the many-worlds interpretation are found in Albert and Loewer (1988), Mensky (2005, 2007a), Squires (1988, 1991), (Zeh 1970), Barrett (1999), and in various contributions to the Oxford University Press Volume *Many worlds?* (ed. by Saunders et al. 2012). The two main issues dealt with in those interpretations are the notion of probability and/or the distribution of consciousness between the parallel worlds.

8. I am also not dealing with the description of the measurement problem via decoherence, since decoherence does not attempt to *explain* the measurement problem – and hence does not suggest an independent interpretation of quantum mechanics.

It appears to be hard to generate clear-cut experimental evidence within physics that can be interpreted in favor of either the Copenhagen interpretation or the many-worlds view.[9] Bohr's complementarity principle, however, closely related to the Copenhagen interpretation and implying that physical entities may either behave as a particle or a wave but never both ways at the same time, got more and more undermined by experimental findings at the double slit.[10] By using very clever experimental designs, some research groups (see, e.g., Mittelstaedt et al. 1987; Scully et al. 1991; Menzel et al. 2012) have demonstrated that it is possible to partially or fully keep the interference pattern (wave-like behavior) whilst nevertheless measuring the path the respective particle was taking. Whereas those findings are unfavorable for the Copenhagen interpretation, they are leaving the many-worlds interpretation untouched.[11] Indeed, in his 1997 popular science book *The fabric of reality*, David Deutsch seems to indirectly base his argument pro many worlds already on those novel findings. When discussing the interesting change of the interference patterns of a singular photon sent through four versus two slits even though the way of the photon through one of the slits can clearly be identified (Deutsch 1997, chapter 2), he leaves the possibility unmentioned which has been demonstrated in an overwhelming number of older experiments: that measuring the path of the photon would often *destroy* (or largely diminish) the interference pattern; and that only very clever experiments lead to the new type of results.

Sure enough, those novel findings at the double slit have not necessarily been interpreted in favor of the many-worlds view by other physicists. Just one, perhaps quite unspectacular example is a poster by Boscá Díaz-Pintado (2007) who discusses, in light of those novel findings, the necessity to change the formalism of quantum mechanics,

9. The situation is unclear enough that David Deutsch and Michael Mensky, two vivid proponents of the many-worlds view on the physics side, disagree on the evidence presented within physics. Deutsch believes that the experimental evidence generated within physics is already in favor of the many-worlds view (Deutsch 1997, chapters 2 and 3). He even *identifies* quantum mechanics with, how he calls it, the *Everett theory* (Deutsch 2012). However, Mensky (2005, 2007a,b, 2010) argues that the evidence generated within physics cannot unambiguously be interpreted in favor of the multiverse view.

10. The original double-slit experiment demonstrating the wave-like nature of light has been carried out first by Young in 1803, already; the first experiment of this type using electrons has been designed and carried out by Jönsson (1961). When carrying out those experiments and measuring the path of the electron (or of other particles) through any of the slits, the interference pattern normally gets destroyed.

11. For another (hypothetical) way of potentially discriminating between different interpretations of quantum mechanics, see Deutsch (1985).

to formulate further assumptions, to modify the complementarity principle etc. Nothing more 'radical' is mentioned.

There are more reasons, however, for taking the many-worlds view seriously (and seeing the Copenhagen interpretation quite critically): Taking the linear Schrödinger equation literally, as a description of the actual world,[12] there is no need or even possibility to think of anything 'artificial' like a state reduction or collapse of the wave function to a singular universe. The Copenhagen interpretation, as convenient as it might be as a workhorse in applied physics, is just not parsimonious. The implied collapse of the wave function is 'alien' to quantum mechanics (Mensky 2005, 2007a, 2010, chapters 1 and 2).[13]

So it looks like if some evidence for the permanent coexistence of parallel worlds already evolves within physics, but skepticism regarding its potential to actually encourage a shift of paradigm towards a many-worlds view (both within and outside physics) is advisable. After all, shifts of paradigms require time and effort (Kuhn [1962] 1996). This is the reason why this article is proposing an interdisciplinary search for evidence for the many-worlds view.

Step 2: Role of consciousness in quantum mechanics – not only important for the multiverse view

Many current theories of consciousness, e.g., many of the approaches proposed in psychology, are characterized by a reductionist approach where the work of consciousness is 'degraded' to some *specific cognitive operations*. Such type of theorizing (as well as the underlying empirical studies) would be categorized as belonging to the 'easy problems' regarding consciousness by David Chalmers (1995, 1996); whereas the 'hard problem' of

12. A more precise view on the Schrödinger equation, accepting human epistemological limits, would be saying that it accurately describes our *room of perceptual possibilities* (see also footnote 38).

13. Neither Deutsch nor Mensky consider it a problem applying the many-worlds perspective to macro phenomena, i.e., our life, whereas the experiments underlying the measurement problem as well as quantum mechanics itself have originally been conducted or developed, respectively, for the world of micro particles. I am sharing this perspective with Deutsch and Mensky. A stream of research that does *not* help with better understanding the micro-macro link, is the experimental work showing that quantum effects (i.e., systems being in superposition states) already occur with somewhat 'larger' objects, with atoms or even molecules (see Venugopalan 2010), or even visible objects under very low temperature (O'Connell et al. 2010). (To keep things clear, only the object, a micromechanical resonator, is visible with the eye, not the quantum effects themselves).

consciousness could be described by questions such as "why are some organisms *subjects* of experience?" or "why do *qualia* exist?"[14]

From the perspective of quantum mechanics, consciousness – in the second, 'hard' interpretation as sort of a *pure subjectivity* – plays a central role in the solution of the measurement problem (e.g., Squires 1988; Mensky 2005, 2007a; Stapp 2009). Quantum systems, including the measurement device etc.,[15] are in a *superposition state* before any conscious observation is conducted; or in other words, a *unique result* or outcome of the measurement is – finally – determined whenever consciousness of the observer comes into play. Thus, paradoxes can be constructed such as Schrödinger's cat or Wigner's friend.[16]

The measurement problem is a fairly complex issue, but even trickier is the question as to how to make use of the effects of conscious observation in a multiverse 'proof.' The reason is that the role that the observer's consciousness plays in the solution of the measurement problem can be interpreted differently, in turn favoring different interpretations of quantum mechanics or being related to changes in its formalism (examples):

1. Consciousness can be seen as the 'force' causing the *collapse of the wave function* hence favoring the Copenhagen interpretation (Stapp 2009).
2. Consciousness might be interpreted as the entity responsible for *separating between infinite 'alternatives'* or *'alternative realities'* (Everett 1957; DeWitt 1970, 1971).
3. Consciousness might also be *identified* with the *selection* of *one* subjective alternative (Squires 1988; Mensky 2005, 2007a,b, 2010).
4. And, based on the *objective reduction formulation* of quantum mechanics, there

14. The term qualia describes the individual's conscious experience and is at the core of the mind-body problem. The term has first been defined in its modern usage by Lewis ([1929] 1956).

15. The view that measurement per se (by whatever device) is at the core of the measurement problem will not be supported here. There is no logical line that can be drawn between either the to be measured physical system and the physical measurement device, between the latter and the perceptual apparatus, between the perceptual apparatus and the brain etc.

16. Wigner's friend is a thought experiment, an extension of the well-known Schrödinger's cat consideration. Here, a friend of the principal investigator conducts a measurement at some quantum system for him, e.g., measures the outcome of Schrödinger's cat experiment, whilst the principal investigator is absent from the laboratory. The question is when the outcome of the measurement is actually determined, only after the return of the principal investigator, or at a previous moment, e.g., when the friend has looked at the device but the principal investigator is not informed about the outcome, yet?

are approaches that link the action of consciousness to *processes in the brain* (e.g., Hameroff and Penrose 1995; Hameroff 2012).

Since nothing can be concluded at this point from the relation between consciousness and the measurement problem for the 'presumptive evidence proof,' the paper will look at the action of consciousness from a different angle, later, when the issue of free will is dealt with. The four exemplary perspectives just presented, however, share the view of consciousness being at the center of a process of 'creation' of subjective reality.

Step 3: Individuals' bodies anticipate the future – and this only makes sense in the multiverse

This subchapter will report on evidence that people are able to anticipate the future. This *fact* is quite important for the multiverse 'proof' to be crafted in this paper because it makes a case against a linear flow of time with important consequences as demonstrated at the end of this subsection. The most conclusive evidence for this ability pertains to *body reactions*. The presented evidence has a close relationship with the findings by Libet and coauthors (e.g., Libet et al. 1982, 1983) as well as Soon and coauthors (Soon et al. 2008) that will play an important role in the next subsection.

In a large-scale meta-study on *anticipatory physiological responses,*[17] Mossbridge et al. (2012) analyze a total of 26 reports published between 1978 and 2010. The authors find strong evidence for individuals' abilities to physiologically anticipate unpredictable events (randomly ordered arousing vs. non-arousing stimuli or guessing tasks with correct/incorrect feedback), no matter what type of physiological measure was used: "electrodermal activity, heart rate, blood volume, pupil dilation, electroencephalographic activity, blood oxygenation level dependent (BOLD) activity" (1). In a fixed effects model, the overall statistical significance for predictive physiological anticipation turned out to be $p < 2.7 \times 10^{-12}$. The evidence is so clear, that (conservatively calculated) 87 unpublished contrary reports would have been necessary to reduce this evidence to chance ($p > 0.05$). Hence it is quite safe to conclude that individuals' bodies are able to anticipate future developments.[18]

17. An example for this type of research is the study by Bierman and Radin (1997) where individuals' electrodermal response significantly differed between emotional and calm pictures already before their presentation.

18. Interestingly, such 'time-backwards' effects have also been proposed as an explanation for some 'strange' behavior of particles (see, e.g., the experimental evidence reported by Herzog et al. 1995).

What does this imply in terms of physical theory? Is there any way of explaining such effects of the future on the present? And why is this evidence supposed to help with the 'existence proof' of parallel realities? The question one wants to ask here is "are there ways to think of time as something that does not just flow in the direction we would normally suppose, from past over present to future?"[19] There are exactly two ways that physics has taken to deal with that question:

1. Scholars have intensively thought about how physical laws could be applied the 'other way around,' i.e., backwards.[20] There is one physical law, however, that seems to contradict such approaches because it appears not to be reversible: the second law of thermodynamics, i.e., the increase of entropy over time. If entropy increases over time, how could we possibly 'go back?' Time-reversing physical laws in a singular world, however, also runs into *logical paradoxes*, described in a graphic way in the form of the 'grandfather paradox'[21] in the literature on time travel. Although time travel seems to be a different pair of shoes than physiological anticipation, *any effects* of anticipating the future potentially leading to changes in an individual's present behavior so that the respective future will *not be reached*, anymore, leads to the same type of paradox.[22]
2. A second, more radical way is to question the idea of a flow of time altogether. Actually, this second approach should be judged as the theoretically convincing way, because it does not run into 'grandfather' type paradoxes. David Deutsch (1991) was the one who introduced a mathematical solution to this problem in his treatment of time travel (see also Deutsch and Lockwood 1994). Time travel does not lead to any logical inconsistencies *if* there are parallel universes.

19. For a systematic analysis of different physical theories regarding our subjectively perceived, asymmetric flow of time see Zeh (1999).

20. A good impression of this type of research can be gained by looking at the numerous theoretical and empirical contributions to *Frontiers of time: Retrocausation – experiment and theory* (ed. by D. P. Sheehan 2006).

21. In the grandfather problem, the time traveler goes back and kills his grandfather at young ages, actually before his father was conceived, so that the time traveler himself should not exist.

22. An important difference between the situation of the grandfather paradox and our situation is that the body of an individual does not have to be 'added' to some reality to change anything there, leading to the problem that mass would have to be either transferred to or to be 'produced' within this reality. For our case it is sufficient that consciousness is able to somehow connect to a reality where a 'replica' of ours already resides (see below).

> Specifically, the logical inconsistency of traveling to and changing one's own past is solved by switching universes. After traveling to the 'past' and 'returning' to 'presence,' the time traveler resides in a new, parallel reality. As already argued above, this consideration of the potential effect of time travel is relevant for our case of anticipatory responses of the human body since changes in the body's reaction that prevent the foreseen future involve the same type of paradox. In his popular science publication *The fabric of reality*, Deutsch (1997, chapter 11) develops different times as special cases of other universes.

Hence, a logically consistent theoretical account for the possibility of predictive physiological anticipation by individuals is only possible on the basis of a permanent coexistence of parallel realities including the coexistence of parallel times. A graphic way of looking at both anticipatory reactions as well as our regular perception of a flow of time might be sort of 'lateral movements' of our consciousness between universes or realities or just 'locations.' Note that this is not to suppose that there is any novel, underlying physics needed for this. The underlying physics is the multiverse.

Interestingly, a *non-presentist* view, i.e., a view where the presence is not seen as the only existing state of the world, cannot solely be developed within the multiverse interpretation of quantum mechanics. The so-called *block universe* view, postulating a four-dimensional world with time as a *permanent* fourth dimension (additional to the three dimensions of space) was already proposed by Minkowski in 1908 (1952, 75) as a consequence[23] of Einstein's special relativity theory. Not surprisingly, then, because of the coexistence of multiple times, Minkowski also postulated the coexistence of multiple spaces:

> We should then have in the world no longer space, but an infinite number of spaces, analogously as there are in three-dimensional space an infinite number of planes. Three-dimensional geometry becomes a chapter in four-dimensional physics. (*Ibid*.)

So the parallel existence of different times can be arrived at from different theoretical starting points.[24] But what about our subjective experience of moving along some linear

23. According to Petkov (2005), the block universe view is the only logically consistent consequence from special relativity.

24. Although the structure of reality derived from special relativity might look quite differently than the one derived from quantum mechanics. It is beyond the scope of this paper to explore those differences in more detail. It is also beyond the scope of this paper to concern itself with relativistic quantum mechanics.

time dimension? The fact that we are normally moving from one reality to another reality, where the second reality is perceived as a 'later' point in time might be seen as a 'convention of conscious experience' or 'perceptual convention,' perhaps rooted in culture. A different perspective on the same phenomenon would be Kant's view of time (Kant [1781] 1996, A30-2/B46-9 and A35-6/B52). According to Kant,[25]

> Time is not an empirical concept that is somehow drawn from experience. For simultaneity or succession would not themselves come into perception if the representation of time did not ground them a priori. Only under its presuppositions can one represent that several things exist at one and the same time (simultaneously) or in different times (successsively). (A30/B46)

The coexistence of different times in parallel realities has an important consequence for the one remaining problem that has been put forward against the potential existence of time-backwards effects: The second law of thermodynamics would not be a problem for 'time-backwards' effects, anymore. In the case of parallel universes, i.e., if 'different times' coexist, different states of physical entities with respect to their entropy would also coexist. And if conscious beings were able to 'laterally move' with their conscious 'emphasis' between those versions of themselves, i.e., across different realities/parallel times, this would imply that they were also able to 'move' between different states of entropy, say, of different versions of their body. That in turn implies that consciousness would be able to also 'move' in the direction of lower entropy,[26] appearing as if the time arrow would have been reversed.

Summarizing this view, our perception of time could be described as taking 'snapshots' of different realities where some 'perceptual convention' or a priori category in the sense of Kant normally organizes them in the form of a unidirectional flow of time (for inspiring empirical findings on this matter varying the duration between 'snapshots' of various courses of action see, e.g., Gruber and Block 2012).[27] Since in principle other

25. Note that not only Kant offers a 'non-objective' account of time. Most idealist philosophers would agree with this basic notion. Of special importance for a 'non-objective' account of time are the thoughts by Leibniz (see, e.g., Grosholz 2011, 347–349), especially well articulated in the Leibniz-Clarke correspondence (Leibniz and Clarke [1717] 2000).

26. A similar line of reasoning is presented by Mensky (2010) to better understand the survival of living beings. He relates this to the 'anthropic principle.'

27. Hameroff (2012) gives an overview of different approaches and results underlining this idea.

points in time are always present, however, this opens the door for time-backwards effects.

Step 4: Free will can only exist in the multiverse

1. Structure of the argument

As already stated in the introduction, empirical results across different cultures demonstrate that most people intuitively believe to possess free will (Nichols 2011). Neuroscience however, seems to prove free will to be an *illusion* (see below). A majority of philosophers has chosen a compatibilist perspective (see, e.g., Dennett 2003), arguing that, under certain conditions, individuals can be held responsible for their actions even if an *actual* free will is absent.[28]

The quite emotional debate about free will and responsibility that took place in the last decades originated in the well-known Libet-experiments (Libet et al. 1982, 1983; Libet 1985) suggesting that the measured readiness potential for a motor action was running ahead of the reported conscious decision. Whereas there has been a critical debate about how to interpret those findings, e.g., by John Eccles (1985),[29] most interpreted them as evidence for (a) free will being *impossible* and (b) subjective perception of possessing free will being an *illusion*.[30]

The debate regained its vigor quite recently with technically more advanced neuroscience studies (Soon et al. 2008) where consciousness not only has been demonstrated to run *several seconds* after specific activities in the brain. But allowing subjects to actually *choose between two alternatives* (i.e., pressing a left or a right key), the authors were able to *predict* the respondents' choice for one of the alternatives based

28. Important other positions are different versions of incompatibilism denying the existence of responsibility under conditions of determinism. Another major position is libertarianism. The most well-known current libertarian is Kane (2003), building his argument pro free will on indeterminism consistent with the Copenhagen interpretation of quantum mechanics. It is unfortunately beyond the scope of this paper to provide a more thorough description and deeper analysis of those important perspectives. A detailed introduction to and discussion of different historical and contemporary perspectives on free will is provided in Walter (2001).

29. The question *how* consciousness might influence (material) brain activities is further analyzed by Beck and Eccles (1992).

30. Since the observed order of events in the experiments is: (1) readiness potential, (2) conscious decision, (3) action, Libet (1999) argued that consciousness might still be able to veto behavior. However, this argument has also been criticized. See, e.g., Velmans (2003) and Kühn and Brass (2009).

on *specific brain areas* that were activated before the conscious decision was reported. Or in other words, when a certain brain area would be activated, consciousness would make a choice for, say, left, a few seconds later, and after that the person would press the left key. The same would hold for the decision to press the right key, but with a *different brain area* activated ahead of time. So the fact that people think they are consciously deciding in favor of pressing a left or a right key simply must be an illusion, no? So how realistic is our perception of free voluntary acts?

In the following it will be argued that Libet's as well as the more recent neuroscience findings can actually be used to *justify* the permanent coexistence of parallel realities. A couple of introductory thoughts are necessary at this point:

- The paper is going to employ a teleological argument.[31] It will be argued that possessing an actual free will gives our consciously experienced life, i.e., qualia, a purpose or meaning.[32]
- The paper will then elaborate on why experiments of the Libet type and modern followers (e.g., Libet et al. 1982, 1983; Libet 1985; Soon et al. 2008) do not necessarily rule out the possibility of an actual free will in the sense of being able to choose A instead of B under identical internal and external causes.
- Later in this subchapter, it will then be discussed what interpretation of the many-worlds view could make free will *possible* and what their respective ontological consequences are. For this means, the paper will first briefly discuss how plausible the many-worlds interpretations are that have already been suggested by Everett-DeWitt, Albert and Loewer (1988), Squires (1988, 1991) as well as Mensky (2005, 2007a,b).[33] A novel interpretation that is free will friendly and ontologically more appealing than the previously suggested ones will also be proposed.

31. "Questions about teleology have, broadly, to do with whether a thing has a purpose or is acting for the sake of purpose, and if so, what that purpose is" (Woodfield [1976] 2010, 1). Teleological or so-called design arguments have, e.g., been crafted in favor of the existence of God (e.g., Aristotle [350 B.C.] 1999, 5–6; Plato [360 B.C.] 2000, Timaeus 28a-34b; Aquinas [1265–1273] 2006, 19) or to disapprove philosophical positions such as the solipsism (Kant [1781] 1996, B 39 et passim).

32. Dennett (1991) tackles the problem in a radically different way by arguing that qualia does not exist, a perspective that will not be followed, here.

33. Recent 'realist' perspectives (see Saunders et al., 2012) will not be discussed.

- Still on the way of completing the argument, the subsequent chapter will deal with the way *how* consciousness might freely choose between alternatives in the multiverse.

2. Free will might not be an illusion if different times are parallel

Regarding the existence of consciousness, a teleological perspective (see footnote 31) might lead to the following question: What could be the 'reason,' the 'sense' of being conscious in the basic meaning of *qualia* (the 'hard-problem' aspect of consciousness; Chalmers 1995, 1996), if there is not any effect of this basic feature of consciousness on our decisions whatsoever? Note that asking this question is inspired by two (related) convictions: (a) Consciousness is not a byproduct of physiological (brain) activity, because qualia, i.e., our conscious experience of life, are something *qualitatively* different from physiological processes.[34] (b) Consciousness is neither supervenient on the physical nor does it influence any physical processes. This is a radical departure from many well-known approaches (e.g., Lewis 1994), that, however, will become more transparent towards the end of this contribution.

Contemplating the question on the 'meaning' of consciousness, one is indeed tempted to conclude that consciousness might have the 'sense' of 'producing' something like a free will. Especially since the alternative perspective on *subjective experience*, watching of and acting in (with fixed roles) a technically advanced 3-D movie, with no possibility to change anything we see, is a view with hardly any teleological appeal.

But then, one might argue: "Nice thought, but how to rule out the argument put forward based on Libet's and followers' experiments? If consciousness is always running after the fact, free will simply *must* be an illusion, no?"

Here is my argument: The discussion in step 3 of my 'proof' lead to the impression that parallel realities might grant us (i.e., our consciousness) with the possibility of *laterally* moving between different times (because they coexist); this also being a theoretically consistent explanation for predictive physiological anticipation by individuals (or time-backwards phenomena in general) that does not run into paradoxes. Assuming the appropriateness of this explanation, however, it is only a small step to also assume that consciousness is able to make backwards-directed decisions, e.g., choose in favor of some motor actions 'backwards' – or better laterally – in time. This in turn would allow for a very different perspective on the Libet type experiments: The fact that

34. A detailed discussion of this important and controversial matter as well as an overview of the relevant literature beyond the 'hard-problem' analysis by Chalmers (1996) is not possible in this paper.

the experience of a conscious decision takes place *after* building the readiness potential for a motor action, or *after* observable activities in certain brain areas, would become meaningless for the free will debate.

3. Analyzing free-will friendliness and ontological consequences of different versions of the multiverse view

In this subchapter, some fundamental versions of the multiverse interpretation will be dealt with. They will all be analyzed regarding their free-will friendliness as well as their ontological consequences. A basic problem pertaining to all those multiverse versions is the question how to deal with the Born rule. Therefore the subchapter starts with this generic problem.

The problem with the Born rule: The Born (1926) rule, successfully used in practical applications of quantum mechanics for many decades and integral part of the Copenhagen interpretation provides specific probabilities for different measurement outcomes. For a multiverse perspective, this causes trouble in two regards: (1) How could one make any sense of probabilities in the multiverse, when in fact the Schrödinger equation is deterministic? How could the Born rule be derived within this framework?[35] (2) How could an actual free will possibly be established if probabilities of measurement appear to be governed by the Born rule?

1. The problem starts with the fact that it is generally unclear (also outside the multiverse view; see, e.g., Landsman, 2008) what exactly justifies the Born rule theoretically (empirically, its support is excellent). After decades of different approaches, a few scholars have quite recently pursued ways to derive the Born rule from *subjective* principles, either decisions (Everettian view: Deutsch 1999; Wallace 2012) or generalized probability theory (quantum bayesianism: e.g., Fuchs 2010). Both approaches assume the application of certain normative principles or axioms.
2. Since an individual may not necessarily be *obliged* to obey to either the rationality axioms proposed by Deutsch (1999) and Wallace (2012) or the generalized probability theory proposed within quantum bayesianism,[36] those

35. See also the discussion in Squires (1991).

36. Outside quantum mechanics, e.g., in economics and psychology, there are large research fields devoted to the understanding of deviations of people from rational decision principles (e.g., Kahneman and Tversky, 1979) or the Bayes rule.

approaches do principally open the space for free will; if, as assumed above, consciousness is not supervenient on the physical. But, given the excellent empirical support for the Born rule, is there actually any *room* for free will? The problem we seem to be facing here arises from a conflict between subjective and intersubjective perception.[37] Measurements carried out in physics as well as psychology laboratories are *reported* and communicated (that's the main point of carrying out scientific research in the first place); their results become intersubjective facts. The Born rule is such an intersubjective fact. If individuals' consciousness would *measurably* and *intersubjectively communicable* influence the observation probability of quantum outcomes in a straightforward and replicable way, this intersubjective fact would be violated. Instead, an individual's influence on developments might rather be expected regarding non-measurable, non-reported, fuzzy, and complex developments; or, in other words: with respect to the individual, personal or better *subjective experience of life*. E.g., meeting the perfect person to marry, as improbable that might objectively be, may (a) nevertheless happen and (b) never violate the Born rule because it can simply not be analyzed within its framework. Admitted, this poses some problems for a direct 'proof' of the existence of free will. This is not saying that it precludes clever experiments on this matter to be carried out in the future. But it helps understanding why evidence does not exist so far and why the existence of free will can only be suggested indirectly at this point, as is the case with the 'proof' of the multiverse (see again the introduction, especially Figure 1, for the underlying logics). The solution to the two problems is intertwined.

Opening the space for free will: EEC framework as a starting point: According to Mensky's (2005, 2007a) multiverse interpretation, the extended Everett concept (EEC), consciousness is indeed able to influence subjective probabilities so that preferred developments of the world are *perceived* with higher probability within the individual's subjective experience, but without changing anything in the *wavefunction*[38] (see also

37. For a related perspective see Mensky (e.g., 2005, 2007a; 2010).

38. Mensky (along with many others) would call the Schrödinger equation the 'objective wavefunction' associating the Schrödinger equation with the physical world (see also footnote 12). An important question is, however, whether or not the wavefunction is really objective. The Schrödinger equation might alternatively be seen as describing accurately our room of perceptual possibilities; close to 'objective' reality, but not identical with it. Since it contains a time dimension and individuals normally organize reality along the time dimension, the setup of the Schrödinger equation exhibits features one would expect from a

the quite similar thought presented in Squires (1988, 18)). This feature arises from the fact that in the EEC interpretation of the multiverse, consciousness is *associated* with the selection of alternatives, a different idea than 'consciousness separating between alternative realities' – the original Everett-DeWitt view. In EEC, consciousness, instead of passively residing with all possibilities given by the Schrödinger equation, gets an active role. According to Mensky, the question of free will can then be addressed as follows: "What is *free will*? … all alternative behavior scenarios are present as superposition components but the subject can compare them with each other and increase the observation probabilities for the alternatives that seem more attractive to her" (Mensky 2007a, 403).

It is quite clear that the EEC interpretation of the multiverse is free will friendly since the individual is supposed to have an influence on what world of the infinite number of worlds to experience: Consciousness is not obliged to 'stay' with all parallel worlds. However, there are three issues with Mensky's concept of free choices that require clarification:

- One issue is that Mensky only 'allows' the unconscious to have access to parallel realities (see, e.g., Mensky 2007b, 2010), a thought consistent with the fact that the best evidence for individuals getting knowledge of the future is physiological (hence unconscious) (Mossbridge et al. 2012); but how could consciousness then make any (free) choices if there is only one reality left to perceive? A potential solution would be that the number of parallel realities that consciousness considers is smaller than the number considered by the unconscious, but sometimes larger than one.[39] Conscious choices between alternatives could subjectively be experienced in the form of phantasies or 'case studies.'[40]
- The other issue is that Mensky's concept somehow equates perception with choice, a problem that will be addressed in the next chapter because sorting this

manmade theory. Certainly, people in different areas of the planet will all get support for the Schrödinger equation. But given the epistemological limits of mankind, the Schrödinger equation might rather be called *intersubjective* than objective.

39. 'Sometimes' is an appropriate description since in many cases choices are made by the unconscious leaving nothing left to decide for consciousness.

40. More precise than the English 'case studies' would be the German term 'Probehandeln' that had already been used by Sigmund Freud.

out is also relevant for the concept of densely and sparsely populated universes, i.e., the novel multiverse version that will be proposed, below.[41]

- Finally, a major problem of the EEC – not directly related to the free-will problem – that will turn out, however, to be quite relevant for the development to be pursued here pertains to the *solipsism*[42] that Mensky's approach necessarily generates. This implies that EEC is ontologically problematic as will be demonstrated in the following.

Towards a free-will-friendly *and* ontologically convincing multiverse interpretation: Different authors (Everett 1957; DeWitt 1970, 1971; Albert and Loewer 1988; Squires 1988, 1991; Mensky 2005, 2007a,b; Zeh 1970) have proposed different basic interpretations of the multiverse.[43] Each of those interpretations offers a different idea about how consciousness is *distributed* between parallel realities. Whereas the EEC concept is accommodating to free will, other existing interpretations are not.[44] The analyzed concepts are somewhat 'strange,' ontologically. Hence, a new multiverse interpretation will tentatively be sketched.

Everett-DeWitt interpretation: The original account by Everett-DeWitt simply postulates that consciousness is *separating* between different realities; those realities being the result of infinite branchings of the universe. This first theory of the multiverse has been criticized by Albert and Loewer (1988). They argue that this approach is incompatible with the conservation of mass problem.[45] Even more critical for the line of arguments presented here, this approach appears to open no room for free choices since branchings are assumed to be 'automatic,' and consciousness is assumed to follow all of them on equal footing.

EEC interpretation: We have seen that the *EEC interpretation is free will friendly* (Mensky 2005, 2007a,b). But EEC has a huge disadvantage, ontologically. To illustrate

41. In psychology, perception and choices are traditionally treated as separate processes (see, e.g., the textbooks by Hayes 1994; Lefton 1994).

42. According to the philosophical position of solipsism, a person can only be sure of her own existence. A nice overview is given by Fumerton (2006).

43. Further interesting interpretations of the multiverse that are, however, not useful in the course of my argument, can be found in Saunders et al. (2012).

44. This also applies to recent 'realist' interpretations of the multiverse that, from my point of view, do not allow for the existence of an actual free will (for an overview of such approaches see Saunders et al., 2012).

45. It is beyond the scope of this article to evaluate this criticism.

this, I am going to provide a simple choice example. For the sake of simplicity, I will not pay any attention to the blurred boundary between choice and perception at this point; this problem will be addressed in the next chapter. A couple, Tim and Louise, jointly decides whether to buy a Volkswagen or a Toyota as the sole family car. Louise wants a Volkswagen; however Tim wishes to buy a Toyota. Let me further suppose that *both* are *fully* successful in perceiving those realities they would like to see (Mensky 2005, 2007a,b, 2010, chapters 1 and 2). So Tim's consciousness realizes a Toyota, Louise's realizes a Volkswagen. This implies having to deal with *two parallel worlds* where in one of them, Louise is happy with Tim and the Volkswagen, whereas in the other, Tim enjoys his marriage with Louise and their Toyota. The problem with this 'wonderful world,' however, can be derived from Table 1 where the two individuals are listed in the rows, the two different realities in the columns.

		Alternative realities	
		Reality 1: VW	*Reality 2: Toyota*
Alternative individuals	*Louise*	Consciousness present	Consciousness absent
	Tim	Consciousness absent	Consciousness present

Table 1: EEC and the 'zombie' problem

The consequence is that there is no alternative reality where *both* individuals are present with their consciousness. From now on, each of the two partners lives with a 'zombie,' since consciousness is turned away crosswise from the respective realities of the spouses. In this example, free will would be rather unlimited, but would have an extremely high price, too: to basically live *alone*. This potential problem of some multiverse interpretations has already been detected by others. Barrett (1999, 186–192) calls it the 'mindless-hulk' problem, and although not crafted for the criticism of EEC (because Barrett's monograph preceded EEC) it fully applies to it.

I would like to again argue here in a teleological sense, by stating that living in a world of 'zombies' would intuitively not make much sense to me and would at least be perceived as quite unappealing or just 'strange' also by many other people. Although

there are well-known proponents of (moderate) *solipsism* such as Schopenhauer, stating that "THE world is my representation" (Schopenhauer [1818] 2010, 23), Kant, e.g., has argued against such a position, actually in form of sort of a teleology: "It still remains a scandal to philosophy and to the general human reason to be obliged to assume, as an article of mere belief, the existence of things external to ourselves … and not to be able to oppose a satisfactory proof to anyone who may call it in question" (Kant [1781] 1996, B 39).[46]

Universal consciousness interpretation: Squires (1988), when suggesting the same kind of 'selection' of one reality by the individual as Mensky (2005, 2007a,b), realized the solipsism problem and also argues in a teleological way: "… how do we ensure that different observers see the same result? … I suppose I am here making the untestable (?) assumption that most people that I meet are conscious" (Squires 1988, 18). But then he makes a radical proposal that must be seen as an independent interpretation of the multiverse:

> The only solution to this problem seems to be that "consciousness" has a unity, i.e., there is, in some sense, one consciousness which knows the result as soon as I … have made an observation. This universal consciousness must then guide the selection of any subsequent observer. (*Ibid*.)

Requiring 'one consciousness' coordinating all individuals' measurements on one consistent picture of the world (Squires 1988, 1991), however, is bringing back a singular reality 'through the backdoor.' Also, this view is *not* free will friendly, since the 'one consciousness' would have to kind of 'dictate' the individuals' measurements/choices.

Many-minds interpretation: Albert and Loewer (1988) propose a 'many-minds view,' related to the earlier one by Zeh (1970).[47] This perspective is closer to the original Everett-DeWitt formulation than the perspectives suggested by Mensky and Squires; other than Everett-DeWitt, however, it explicitly brings in a probabilistic element. Albert and Loewer (1988) propose an infinite number of minds whose proportions of perceiving one or the other outcome of a measurement are assumed to resemble the probabilities

46. The following humorous statement by Karl Popper shows how difficult this discussion actually is: "I know that I have not created Bach's music or Mozart's…I just do not have it in me" (Popper [1956] 1999, 83). Although this consideration nicely demonstrates that Popper simply cannot be *alone*, it does not necessarily lend support to other visible entities possessing consciousness.

47. Differences between those authors' and Zeh's (1970) 'many minds view' will not be analyzed in this article.

of the "experimentally verified probability rule of quantum theory" (i.e., the Born (1926) rule; see also Squires 1991, 283, in an article comparing his and Albert and Loewers' (1988) view). So if two outcomes of a measurement are, say, equally probable, half of the minds will see one of the two outcomes, and the other half will see the alternative outcome. The authors admit that "this talk of infinitely many minds sounds *crazy*" (Albert and Loewers 1988, 207);[48] Squires (1991) adds that he is not sure "… that the idea of an infinite number of existing minds … makes ontological sense" (285). Since the probabilities are assumed to be given, Albert and Loewer's interpretation is not free will friendly, either.

Densely and sparsely populated universes: So we are left with two equally problematic alternatives; the free-will-friendly EEC by Mensky, leading to solipsism, and all other interpretations not being free will friendly for different reasons. At the core of the problem is the question how consciousness is assumed to be *distributed* between alternative realities. All interpretations that have been proposed, so far, served the extremes: Consciousness is seen as residing with just one or all realities.

But what is the alternative? One possibility would be having densely and sparsely 'populated' universes in terms of the amount of consciousness allocated to them.[49] Let me introduce this concept by using the allegory of a torch light,[50] whose cone of light is brightest in the middle, and where the light intensity fades with more and more distance from the center. Let me assume that each individual's consciousness is distributed in the same way as this cone of light. There is one reality where the center of consciousness resides, and there are neighboring realities where less consciousness resides. The 'distance' from the center is measured in terms of differing choices. Let us look at a situation where our individual in the middle of the cone of light (the one where the center of consciousness resides) decides to take a left turn at some traffic light using her car. In the multiverse, there will always be a 'replica' (a term used by Deutsch in many of his publications) taking the right turn. Now, the 'replica' taking the right turn is slightly off the center, with slightly reduced consciousness. The more the choices of a certain 'replica' differ from the choices of the 'center individual,' i.e., the larger the distance from it, the less bright the light of the cone, and consequently, the lower the amount

48. Sure enough, they developed this account for one purpose, only: to solve theoretical problems of the Everett-DeWitt formulation that they had earlier discussed in their article.

49. I am very grateful to Tanja Strohm for suggesting this solution to me in a discussion.

50. This allegory has the highest intuitive appeal with an old-fashioned torch light, since LED and laser have a more concentrated cone of light.

of consciousness allocated to this 'replica.' In other words, there is a smooth removing of consciousness from realities that are close to the 'center individual,' a strong removing of consciousness, however, from those that are located 'many decisions away.'[51] Note that a similar allegory has been used in the philosophy of time: the 'moving spotlight.' "According to the moving spotlight theory of time, the property of being present moves from earlier times to later times, like a spotlight shone on spacetime by God" (Skow 2012, 223). However, although the 'moving spotlight' theory assumes the parallel existence of different times, there are important differences to the concept presented here. Here, not only times are parallel but also alternative realities at each point in time that are separated by decisions. Also, each individual (including her 'replicas') is using a separate torch light, whereas the moving spotlight is assumed to be 'universal' leading to an *absolute* past, presence, and future (Skow 2009, 2012). Finally, whereas an ideal spotlight has sharp boundaries and shines on just one time, the torch light in our allegory shines on many realities, albeit with diminishing intensity with higher distance from the center.

Given this reasoning, we may either find ourselves in rather densely populated universes, defined, say, as a cluster of 'similar realities,' where a lot of consciousness from many individuals resides (where many bright areas of the light cones meet); the condition being that many individuals have made decisions that get them into those 'similar reality clusters.' Or we are going to find ourselves in sparsely populated universes, where only few people have made choices leading them into our reality, and, consequently, where consciousness of others is involved to a smaller degree; and there might certainly be many cases where the situation is located somewhere in the middle between those two possibilities. However, since consciousness is only removed smoothly, there are no universes with actual 'zombies.' Or to stay within the allegory of the torch light, there is no darkness around the individual, even if only distant parts of the light cones of the other individuals/'replicas' reach that reality.

Although this novel interpretation clearly needs to be further elaborated in future contributions, I would like to argue that it is *free will friendly* because people have an influence on the reality to be experienced (with what degree of consciousness) and that it *makes more sense, ontologically*, than interpretations leading to either solipsism or to

51. Clearly, two questions are open to debate. First, it is unclear whether the 'center individual' will always be perfect in 'picking' the reality that is 'best' for her life or survival (and only the 'replicas' are characterized by less optimal choices). In fact, this might be very unlikely in case of, e.g., unresolved traumata or auto-destructive motives. Second, it is unclear whether only the 'center individual' has the power to make choices (and drags the others along), or whether each of the 'replicas' has some (perhaps small) influence on where the light cone moves (making the presented concept slightly more complicated).

consciousness splitting according to the Born rule etc. Following this novel interpretation, consciousness is *partially* decoupled from the physical world by being able to choose how much emphasis to put on what types of realities.

4. Objective reduction and entanglement as an alternative to 'save free will?'

Quite recently, Hameroff (2012) also argued that Libet-type findings might be consistent with free will if consciousness where able to influence the actions of the brain/body as well as individual's choices 'backwards'; and, based on quantum brain biology, he is convinced that consciousness has this capability. However, there are two reasons why one might question that Hameroff's theory, based on the *objective reduction modification* of quantum mechanics, is able to 'save free will':

1. *Reappearance of all paradoxes connected to time-backwards effects*: Hameroff's argument that only 'acausal' information will be sent backwards (Hameroff 2012, 11) is hard to swallow. Either the respective information *changes something*, e.g., a choice, or it doesn't, where in the latter case it is irrelevant, no? Only the *multiverse* interpretation of quantum mechanics is able to account for changes in the 'past' that are inspired by the 'future' and in turn change the 'future' (see Deutsch 1991). I simply do not see how any single-universe interpretation or the objective reduction formalism – both involving some sort of collapse of the wave function – would allow for this.
2. *The material world has to wait for all of us?* If one follows Hameroff's theory regarding the fact that microtubules in the brain are able to maintain quantum states for a substantial time period (recent evidence appears to be in favor of this part of his theory; Science Daily, January 16th, 2014), how would free will play out outside the respective individuals' brain in a singular universe? Does the outside world 'wait' for, say, one or two seconds for *each* individual's brain to decide what reality to 'select,' and how would 'bargaining' between different brains take place if preferences are different?[52]

52. Note that there is a similarity between this 'bargaining requirement' and the argument made by Squires (1988) in the framework of his multiverse interpretation: universal consciousness; that perspective has already been critically discussed above.

Those arguments show that Hameroff's (2012) way of demonstrating the possibility of free will in an objective reduction framework is implausible and hence no alternative to the respective claim based on the multiverse interpretation presented here.

How does free will act in the multiverse?

It turned out that free choices appear to be in principle possible if certain versions of the multiverse interpretation are adopted (either EEC with the unappealing consequence of solipsism or the densely and sparsely populated universes interpretation). So it might be tempting to ask how that works. Consciousness is associated with *perceiving* a specific outcome of the measurement process (see above). In psychology, perception and choices are traditionally treated as separate processes (see, e.g., the textbooks by Hayes 1994; Lefton 1994). So *how* could consciousness actually produce *free choices*?

One possibility of interpreting the action of consciousness is indeed that an individual's choices are *automatic*, given the perception that she has, and that free will works *indirectly*, via the ability of consciousness to influence what will actually be perceived. This is a complex thought, and an example will be used to clarify. It starts with a classical (non-quantum), decision-theoretic analysis: Julia wants to buy either a Volkswagen or a Toyota. If she perceives the Volkswagen as more reliable than the Toyota, she will buy it (unlike in the above example with Louise and Tim, there is no conflict between partners here; we may think of Julia being a single). If she perceives the Toyota as more reliable, she will buy that car. Thus, given her *preferences* (only reliability is relevant!) and her respective (automatic) *perception* of the reliability of the two cars her choice is fully determined. This simplified decision-theoretic analysis is depicted in Figure 2.[53] In this as well as in the subsequent Figures 3 and 4, the smiley represents the point where people think they decide.

53. For the sake of simplicity, the analysis is abstracting here from many complexities of those decisions, i.e., using heuristics, falling prey to biases etc. This picture is hence closer to a normative rather than descriptive decision-theoretic account.

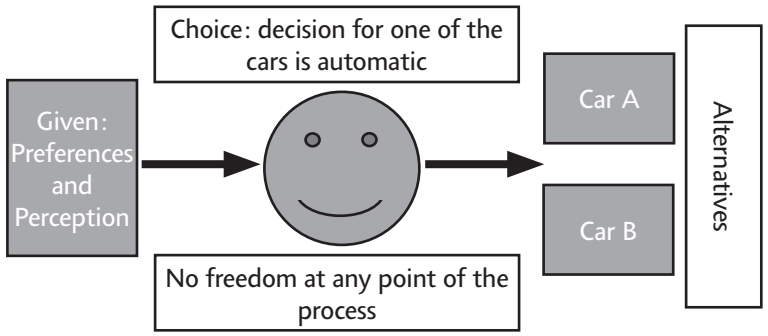


Figure 2: Free will in a simplified decision-theoretic framework

But let us now assume that her consciousness is able to *choose how she perceives* the reliability of those two cars simply by 'choosing' that alternative reality (more precisely, influencing the probability of subjective observation) in which one or the other car *is* more reliable.[54] Then, free will could play out in the *choice of the reliability perception* or more precisely, in enlarging the probability to observe this specific reality; given this operation, the choice of the car is still automatic (see Figure 3).

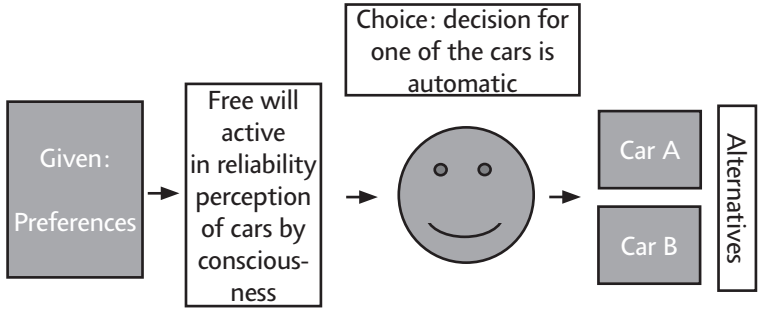


Figure 3: Free will when choosing how to perceive different realities

54. This idea (depicted in Figure 3) as well as the following idea (depicted in Figure 4) are somewhat inspired by Mensky's idea of postcorrection (Mensky 2007b). Specifically, the processes depicted in Figures 3 and 4 are two possible interpretations of the process of postcorrection, based on a more explicit differentiation between perception and choice than in Mensky's treatment.

Another possibility of thinking about this problem is that not a certain characteristic of an object is perceived (here, the reliability of a car), and a decision will automatically be made based on this perception, but instead the 'attractiveness' of different alternative realities already *including the choice* of a specific car (see Figure 4). Let us assume that parts of those possible realities are the different choices that Julia *has made*. So Julia can 'opt' between perceiving a reality in which she *has chosen* a Volkswagen and a reality where she *has chosen* a Toyota. If the reality with the Volkswagen turns out to be more attractive (still with the reliability of the car being the only component that differs), her consciousness may opt for perceiving this version as 'real.' It is important to note how consistent this description appears to be with what was discussed above as reinterpretation of the findings by Libet and coauthors: It was argued that consciousness might be able to work backwards. Indeed, this might be the way how choices are generally made. The interesting aspect here, however, is that our intuition as to what a choice is turns out to be somewhat violated; perception or *choices what to perceive*, might be sufficient: Everything might be about choices what reality to perceive!

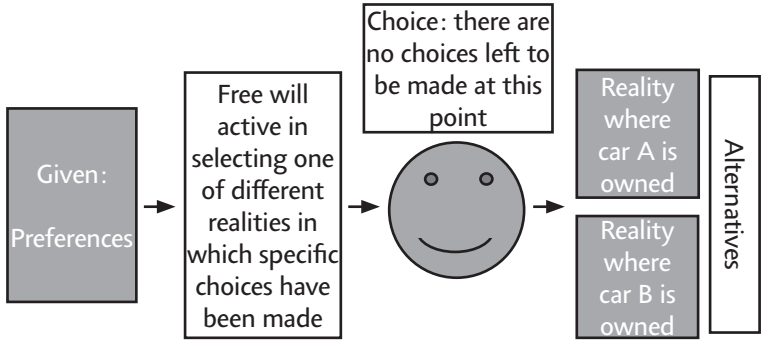


Figure 4: Free will when choosing in which reality to reside

Since the theoretical perspective depicted in Figure 4 appears to be the one that is most consistent with the reasoning presented in other parts of this paper, it makes sense to look at the consequences it might have on the sort of free will that the multiverse grants us with: It is a *freedom of perception*. We have the perceptual freedom to opt for experiencing certain realities rather than others. Moreover, this means that although consciousness is not supervenient on the physical, at the same time is has no influence on the physical (such a claim was earlier made and it should have become more transparent

at this point). But although consciousness has no influence on the physical, possibilities of perception are in principle infinite – even if the *degree* of flexibility in actually choosing between those 'films' is unclear at this point and might differ considerably between individuals and situations.

Conclusion and remarks

The results of the presented analysis are fourfold. First, it could be shown that the existence of free will critically depends on the existence of parallel realities; those multiple realities are characterized by the coexistence of different times as well as the coexistence of different, decision-dependent versions (i.e., 'replicas') of each individual (or perhaps better: parallel 'worlds'). Second, since free will appears to be only possible in the multiverse, a teleological argument claiming that qualia should have a purpose and that free will should be that purpose leads in turn to a preference for the multiverse interpretation over other interpretations of quantum mechanics. Third, the paper proposes a novel form of the multiverse interpretation that is not only free will friendly but might also be seen as ontologically acceptable. Fourth, the paper analyzes how free choices might be made in the multiverse and arrives at the conclusion that free will is about choices what reality to perceive.

To craft the 'presumptive evidence proof,' this contribution had to deal with theories and empirical findings from quite a few different areas, and it is beyond my expertise in many of those areas to formulate precise future research opportunities. Indeed, readers who are experts in the respective areas will certainly find it much easier to develop such ideas. Therefore I am taking the liberty, here, to develop a rather personal perspective by asking the question what I would like to work on next, what I perceive as exciting paths for future research, and to select two questions that I feel are of most interest to the readers of this journal.

As a researcher originating from the decision sciences, the first thing that comes into my mind is the large impact that the thoughts presented in this paper might have on the decision sciences as well as game theory. Very recent research labelled "quantum social science" (Haven and Khrennikov 2013) has already taken the route of linking quantum mechanics and questions in the decision sciences, but in a way radically different from what would result from the analysis presented in this paper. Specifically, the analysis by Haven and Khrennikov (2013) avoids the multiverse interpretation of quantum mechanics and just applies the quantum formalism to human decision making whilst also avoiding to answer the question why it should be relevant to it:

> We emphasize that in our approach, the quantum-like behavior of human beings is not a consequence of quantum-physical processes in the brain. Our premise is that information processing by complex social systems can be described by the mathematical apparatus of quantum mechanics. (Haven and Khrennikov 2013, xviii)

Whereas the analysis presented in this paper would require also linking the 'content,' a deeper understanding of the multiverse interpretation of quantum mechanics to the actions of consciousness in selecting between alternative realities. It would be most appropriate to collaborate on this matter with a quantum physicist open for (or convinced of) the multiverse interpretation of quantum mechanics as well as open for its application to human decision making.

Another interesting route to be taken might be the analysis of the consequences of the presented approach to philosophy in general, over and above the free-will question. For instance, does the multiverse interpretation of quantum mechanics offer a novel scientific underpinning for philosophical idealism? Would the multiverse view fit best with the early Hindu/Vedanta version of idealism, or would it rather fit with Neo-Platonism, with the philosophies by Spinoza or Hegel, or with German idealism? Here, collaboration with a philosopher knowledgeable in different idealistic philosophy traditions would be most suitable.

Finally, what are the consequences of the multiverse view on how we see and live our lives? The most dramatic shift when moving towards a multiverse view might be the resulting understanding of the plethora of possibilities how to experience our life since free will, most probably in the version of a freedom of perceived choice, can be justified within this framework. Even though the exact flexibility we would have for perceiving different realities is unclear and most certainly dependent of the individual and her circumstances, this might, in turn, leave not much space for people feeling as 'victims of circumstances.' Rather, it should strengthen the perception of responsibility. And people who understand having an actual influence on what they experience in their lives might also act differently, less fearful, perhaps, and more optimistic.

References

Albert, David and Barry Loewer. 1988. "Interpreting the many worlds interpretation." *Synthese* 77 (2): 195–213.

Aquinas, Saint Thomas. (1265–1273) 2006. "Summa theologiae: God's will and providence." In Vol. 5 of *The dominican council*, edited by Thomas Gilby, 19–26. Cambridge: Cambridge University Press.

Aristotle. (~ 350 B.C.) 1999. "Physics, Book VIII." In *Aristotle physics,* edited by Daniel W. Graham. Oxford: Clarendon Press.

Auletta, Gennaro. 2001. *Foundations and interpretation of quantum mechanics*. Singapore: World Scientific Publishing Co.

Barrett, Jeffrey A. 1999. *The quantum mechanics of minds and worlds*. Oxford: Oxford University Press.

Beck, Friedrich and John C. Eccles. 1992. "Quantum aspects of brain activity and the role of consciousness." *Proc. Nadl. Acad. Sci. USA* 89: 11357–11361.

Bierman, Dick J. and Dean I. Radin. 1997. "Anomalous anticipatory response on randomized future conditions." *Perceptual and Motor Skills* 84 (2): 689–690.

Born, Max. 1926. "Zur Quantenmechanik der Stoßvorgänge." *Zeitschrift für Physik* 37 (12): 863–867.

Boscá Díaz-Pintado, María C. 2007. "Updating the wave-particle duality." Paper presented at the 15th UK and European Meeting on the Foundations of Physics, Leeds, March 29–31.

Chalmers, David J. 1995. "Facing up to the problem of consciousness." *Journal of Consciousness Studies* 2 (3): 200–219.

Chalmers, David J. 1996. *The conscious mind in search of a fundamental theory*. Oxford: Oxford University Press.

Dennett, Daniel C. 1991. *Consciousness explained*. Boston: Little Brown.

Dennett, Daniel C. 2003. *Freedom evolves*. New York: Viking Press.

Deutsch, David. 1985. "Quantum theory as a universal physical theory." *International Journal of Theoretical Physics* 24 (1): 1–41.

Deutsch, David. 1991. "Quantum mechanics near closed timelike curves." *Physical Review* 44: 3197–3217.

Deutsch, David and Michael Lockwood. 1994. "The quantum physics of time travel." *Scientific American* 270: 68–74.

Deutsch, David. 1997. *The fabric of reality: Towards a theory of everything.* Middlesex: Penguin Books Ltd.

Deutsch, David. 1999. "Quantum theory of probability and decisions." *Proceedings of the Royal Society of London* 455: 3129–3137. doi: 10.1098/rspa.1999.0443.

Deutsch, David. 2012. "Apart from universes." In *Many worlds? Everett, quantum theory, & reality*, edited by Simon Saunders, Jonathan Barrett, Adrian Kent and David Wallace, 542–552. Oxford: Oxford University Press.

DeWitt, Bryce S. 1970. "Quantum mechanics and reality: Could the solution to the dilemma of indeterminism be a universe in which all possible outcomes of an experiment actually occur?" *Physics Today* 23: 30–40.

DeWitt, Bryce S. 1971. "The many-universes interpretation of quantum mechanics." In *Foundations of quantum mechanics,* edited by Bernard D'Espagnat. New York: Academic Press.

Eccles, John C. 1985. "Mental summation: The timing of voluntary intentions by cortical activity." *Behavioral and Brain Sciences* 8 (4): 542–543.

Everett, Hugh, III. 1957. "'Relative state' formulation of quantum mechanics." *Reviews of Modern Physics* 29 (3): 454–462.

Fuchs, Christopher A. 2010. "QBism: The perimeter of quantum baysianism." Available online at http://arxiv.org/abs/1003.5209.

Fumerton, Richard A. 2006. "Solipsism." In *Encyclopedia of Philosophy* Vol. 9, 2nd ed., edited by Donald M. Borchert, 115–122. Detroit: Macmillan Reference USA.

Grosholz, Emily. 2011. "Reference and analysis: The representation of time in Galileo, Newton, and Leibniz." 2010 Arthur O. Lovejoy Lecture. *Journal of the History of Ideas* 72 (3): 333–350.

Gruber, Ronald and Richard A. Block. 2012. "Experimental evidence that the flow of time is a perceptual illusion." Paper presented at the "Toward a Science of Consciousness" conference, Tucson, Arizona, April 9–14.

Hameroff, Stuart R. and Roger Penrose. 1995. "Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness." *Neural Network World* 5: 793-804.

Hameroff, Stuart R. 2012. "How quantum biology can rescue conscious free will." *Frontiers in Integrative Neuroscience* 6, article 93: 1–17. doi: 10.3389/fnint.2012.00093.

Haven, Emmanuel and Andrei Khrennikov. 2013. *Quantum social science*. Cambridge: Cambridge University Press.

Hayes, Nicky. 1994. *Foundations of psychology*. New York: Routledge.

Herzog, Thomas J., Paul G. Kwiat, Harald Weinfurter and Anton Zeilinger. 1995. "Complementarity and the quantum eraser." *Physical Review Letters* 75: 3034–3037.

Jönsson, Claus. 1961. "Elektroneninterferenzen an mehreren künstlich hergestellten Feinspalten." *Zeitschrift für Physik* 161 (4): 454–474.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263–292.

Kane, Robert H. 2003. "Free will: New directions for an ancient problem." In *Free will*, edited by Robert H. Kane, 222–248. Oxford: Blackwell.

Kant, Immanuel. (1781) 1996. *Critique of pure reason*. Unified ed. Indianapolis: Hackett Publishing Company.

Kühn, Simone, and Marcel Brass. 2009. "Retrospective construction of the judgement of free choice." *Consciousness and Cognition* 18 (1): 12–21.

Kuhn, Thomas S. (1962) 1996. *The structure of scientific revolutions*. 3rd ed., Chicago: The University of Chicago Press.

Landsman, Nicholas P. 2008. "The conclusion seems to be that no generally accepted derivation of the Born rule has been given to date, but this does not imply that such a derivation is impossible in principle." In *Compendium of quantum physics*, edited by Friedel Weinert, Klaus Hentschel, Daniel Greenberger and Brigitte Falkenburg. Berlin: Springer-Verlag.

Lefton, Lester A. 1994. *Psychology*. Needham Heights: Allyn & Bacon.

Leibniz, Gottfried W. and Samuel Clarke. (1717) 2000. *Correspondence*. Edited, with Introduction, by Roger Ariew. Indianapolis: Hackett Publishing Co. Inc.

Lewis, Clarence I. (1929) 1956. *Mind and the world order*. Outline of a theory of knowledge. 1st ed. New York: Dover Publications.

Lewis, David. 1994. "Reduction of mind." In *A companion to the philosophy of mind*, edited by Samuel Guttenplan, 412–431. Oxford: Blackwell.

Libet, Benjamin, Elwood W. Wright, Jr. and Curtis A. Gleason. 1982. "Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts." *Electroencephalography and Clinical Neurophysiology* 54 (3): 322–335.

Libet, Benjamin, Curtis A. Gleason, Elwood W. Wright, Jr. and Dennis K. Pearl. 1983. "Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): The unconscious initiation of a freely voluntary act." *Brain* 106: 623–642.

Libet, Benjamin. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action." *The Behavioral and Brain Sciences* 8: 529–539.

Libet, Benjamin. 1999. "Do we have free will?" *Journal of Consciousness Studies* 6 (8–9): 47–57.

Mensky, Michael B. 2005. "Concept of consciousness in the context of quantum mechanics." *Physics – Uspekhi* 48: 389–409.

Mensky, Michael B. 2007a. "Quantum measurements, the phenomenon of life, and time arrow: The great problems of physics (in Ginzburg's terminology) and their interrelation." *Physics – Uspekhi* 50: 397–407.

Mensky, Michael B. 2007b. "Postcorrection and mathematical model of life in Extended Everett's Concept." *NeuroQuantology* 5 (4): 363–376.

Mensky, Michael B. 2010. *Consciousness and quantum mechanics: Life in parallel worlds*. Singapore: World Scientific Publishing Co.

Menzel, Ralf, Dirk Puhlmann, Axel Heuer and Wolfgang P. Schleich. 2012. "Wave-particle dualism and complementarity unraveled by a different mode." *Proceedings of the National Academy of Sciences* 109: 9314–9319. doi: 10.1073/pnas.1201271109.

Minkowski, Hermann. 1952. "Space and time." In *The principle of relativity: A collection of original memoirs on the special and general theory of relativity,* edited by Hendrik A. Lorentz, Albert Einstein, Hermann Minkowski, and Hermann Weyl, 75–91. New York: Dover. (First presented at the 80th meeting of natural scientists, Cologne, Germany, Sept. 21, 1908).

Mittelstaedt, Peter, A. Prieur and R. Schieder. 1987. "Unsharp particle-wave duality in a photon split-beam experiment." *Foundations of Physics* 17 (9): 891–903.

Mossbridge, Julia, Patrizio Tressoldi and Jessica Utts. 2012. "Predictive physiological anticipation preceding seemingly unpredictable stimuli: A meta-analysis." *Frontiers in Psychology* 3: article 390. doi: 10.3389/fpsyg.2012.00390.

Neumann, Johann von. (1932) 1996. *Mathematische Grundlagen der Quantenmechanik*. 2nd ed. Berlin: Springer-Verlag.

Nichols, Shaun. 2011. "Experimental philosophy and the problem of free will." *Science* 331 (6023): 1401–1403.

O'Connell, Aaron D., Max Hofheinz, Markus Ansmann, Radoslaw C. Bialczak, Mike Lenander, Erik Lucero, Matthew Neeley, Daniel Sank et al. 2010. "Quantum ground state and single-phonon control of a mechanical resonator." *Nature* 464: 697–703.

Penrose, Roger. 1994. *Shadows of the mind: A search for the missing science of consciousness*. Oxford: Oxford University Press.

Petkov, Vesselin. 2005. *Relativity and the nature of spacetime*. Berlin: Springer-Verlag.

Plato. (~ 360 B.C.) 2000. *Timaeus*. Translated by Donald J. Zeyl. Indianapolis: Hackett Publishing Company, Inc.

Popper, Karl R. (1956) 1999. *Realism and the aim of science.* 2nd ed. Guildford and King's Lynn: Biddles.

Saunders, Simon, Jonathan Barrett, Adrian Kent, and David Wallace, eds. 2012. *Many worlds? Everett, quantum theory, & reality.* Oxford: Oxford University Press.

Schopenhauer, Arthur. (1818) 2010. "The world as will and representation." In *Schopenhauer: The world as will and representation*, edited by Christopher Janaway, Volume: 1, 1st ed. Cambridge: Cambridge University Press.

Science Daily. 2014. "Discovery of quantum vibrations in 'microtubules' inside brain neurons supports controversial theory of consciousness." January 16th, 2014. Available online at http://www.sciencedaily.com/releases/2014/01/140116085105.htm.

Scully, Marian O., Berthold-Georg Englert and Herbert Walther. 1991. "Quantum optical tests of complementarity." *Nature* 351: 111–116.

Sheehan, Daniel P, ed. 2006. *Frontiers of time: Retrocausation – experiment and theory*. Melville, NY: American Institute of Physics.

Skow, Bradford. 2009. "Relativity and the moving spotlight." *The Journal of Philosophy* 106 (12): 666–678.

Skow, Bradford. 2012. "Why does time pass?" *Noûs* 46 (2): 223–242.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze and John-Dylan Haynes. 2008. "Unconscious determinants of free decisions in the human brain." *Nature Neuroscience* 11 (5): 543–545.

Squires, Euan J. 1988. "The unique world of the Everett version of quantum theory." *Foundations of Physics Letters* 1: 13–20.

Squires, Euan J. 1991. "One mind or many – a note on the Everett interpretation of quantum theory." *Synthese* 89 (2): 283–286.

Stapp, Henry P. 2009. *Mind, matter, and quantum mechanics*. 3rd ed. Berlin: Springer-Verlag.

Velmans, Max. 2003. "Preconscious free will." *Journal of Consciousness Studies* 10 (12): 42–61.

Venugopalan, Anu. 2010. "Quantum interference of molecules: Probing the wave nature of matter." *Resonance: Journal of Science Education* 15 (1): 16–31.

Vollmer, Gerhard. (1975) 2002. *Evolutionäre Erkenntnistheorie* [Evolutionary epistemology]. 8th ed. Leipzig: Hirzel.

Wallace, David. 2012. "How to prove the Born rule." In *Many worlds? Everett, quantum theory, & reality*, edited by Simon Saunders, Jonathan Barrett, Adrian Kent and David Wallace, 227–263. Oxford: Oxford University Press.

Walter, Henrik. 2001. *Neurophilosophy of free will. From libertarian illusions to a concept of natural autonomy.* Translated by Cynthia Klohr. Cambridge: A Bradford Book - The MIT Press.

Woodfield, Andrew. (1976) 2010. *Teleology*. 2nd ed. Cambridge: Cambridge University Press.

Zeh, H. Dieter. 1970. "On the interpretation of measurement in quantum theory." *Foundations of Physics* 1 (1): 69–76.

Zeh, H. Dieter. 1999. *The physical basis of the direction of time*. 3rd ed. Berlin: Springer-Verlag.

Journal of Cognition and Neuroethics

The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative

Jay Spitzley
Georgia State University

Biography
Jay Spitzley is pursuing a Masters of Arts degree in philosophy at Georgia State University. His research interests include moral psychology, action theory, experimental philosophy, and neuroethics. He received his BA at the University of Michigan and can be contacted at jspitzley1@student.gsu.edu.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation
Spitzley, Jay. 2015. "The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative." *Journal of Cognition and Neuroethics* 3 (1): 363–382.

The Importance of Correctly Explaining Intuitions: Why Pereboom's Four-Case Manipulation Argument is Manipulative

Jay Spitzley

Abstract

Recent empirical findings have shown that intuitions are significantly influenced by subtle and seemingly irrelevant factors. In light of these findings, I argue that before making claims about what best explains intuitions regarding thought experiments, one must acknowledge the effects that certain psychological influences have on intuitions. To demonstrate how problematic it can be to ignore these covert factors, I discuss Derk Pereboom's four-case manipulation argument. While Pereboom claims that intuitions regarding his argument for incompatibilism reliably track relevant features of the four cases, I argue instead that these intuitions are likely driven by order effects motivated by unconscious psychological influences and that these order effects put significant pressure on Pereboom's argument.

Keywords

Intuition, moral judgment, order effects, moral responsibility, Derk Pereboom, manipulation

I. Introduction

It has become common in philosophy to use intuitions about thought experiments and hypothetical cases to bolster one's argument. While we like to think intuitions are reliable, recent empirical research shows that this isn't always the case. Experimental philosophers who contribute to the "negative" program of experimental moral philosophy have discovered that intuitions are not universally held.[1] Rather, judgments vary according to ethnicity (Weinberg et al. 2001), gender (Buckwalter and Stich 2011), and linguistic background (Vaesen et al. 2013). Further research shows that intuitions are unreliable in an additional sense. That is, intuitions and moral judgments are significantly influenced by trivial and rationally irrelevant factors of hypothetical cases, such as the order in which information is presented (Weigmann et al. 2012; Schwitzgebel and Cushman 2012), the way in which the information is worded (Petrinovich and O'Neill 1996), the emotional

1. Experimental philosophy's "negative program," generally seeks to challenge the usefulness in appealing to philosophical intuitions as a method of uncovering justified beliefs (Alexander, Mallon, and Weinberg 2014).

status of the reader (King and Hicks 2011; He et al. 2013; Guiseppe et al. 2012), and even the clean smell of Lysol (Tobia et al. 2013).

While the variation of intuitions across demographics has led some to argue that intuitions about these hypothetical cases should not be used as evidence for philosophical views (Weinberg 2008; Sinnott-Armstrong 2008), I will focus on problems that result from our intuitions being unreliable in the other sense. Specifically, I address problems that arise from features and psychological influences we are largely unaware of driving our intuitions and moral judgments. Given the sway such factors have on intuitions about hypothetical cases and thought experiments, I argue one must proceed cautiously when presenting an argument that relies on an explanation for what features of a case motivate intuitions about that case. Furthermore, I argue that failure to consider these psychological influences (some of which may be entirely unconscious) as alternative explanations for what drives intuitions can undermine one's argument.

Despite overwhelming evidence that humans are bad at knowing what influences their judgments (King and Hicks 2011; Mlodinow 2012; Li et al. 2008), and that even philosophers are susceptible to unconscious psychological influences (Schwitzgebel and Cushman 2012; Tobia 2013), philosophers frequently assume they know what drives intuitions. To demonstrate the importance of taking this new evidence into account for philosophical debate, I discuss Derk Pereboom's (2014) four-case manipulation argument. The success of his argument hinges on knowing what motivates intuitions about the four cases he presents the reader. I argue that by neglecting to consider an alternative explanation for what drives intuitions, namely, order effects, Derk Pereboom leaves open a serious objection to his argument.

II. Pereboom's Four-Case Manipulation Argument

In an attempt to demonstrate that the compatibilist conditions for moral responsibility are insufficient on the grounds that determinism, when properly understood, is incompatible with moral responsibility, Derk Pereboom (2014) presents a manipulation argument. Pereboom attempts to show that even in cases when all compatibilist requirements for free will and moral responsibility are met, agents can still lack moral responsibility. To achieve these aims, Pereboom presents four cases.

Each case involves an agent, Plum, who is causally determined by factors beyond his control to kill another agent, White. Additionally, in each case Plum satisfies all purported

compatibilist requirements for free will and moral responsibility.[2] In Case 1, Plum's mental states are manipulated via radio-like technology by a team of neuroscientists in such a way that he reasons egoistically and decides to kill White. In Case 2, Plum is just like an ordinary human being except that neuroscientists manipulate him in the beginning of his life in such a way that he will later reason egoistically and kill White. In Case 3, the training practices of Plum's community causally determine that he reasons egoistically such that he kills White. Last, Pereboom presents Case 4, wherein Plum is an ordinary human being in a deterministic universe and Plum's egoistic decision to kill White is causally determined by the past and laws of nature. Again, in all four cases Plum satisfies *all* purported compatibilist requirements for free will and moral responsibility *and* Plum's actions are causally determined by factors outside of his control.[3] Pereboom claims: "The salient factor that can plausibly explain why Plum is not responsible in all of the cases is that in each he is causally determined by factors beyond his control to decide as he does. This is therefore a sufficient, and I think also the best, explanation for his non-responsibility in all of the cases" (2014, 79).

Given this presentation, whether Pereboom's argument successfully poses a problem for compatibilist accounts of free will and moral responsibility depends on a few conditions being met. First, readers must not be confused about the causal nature of determinism. Second, readers must truly understand that Plum meets all compatibilist requirements for moral responsibility. Third, readers must find Plum intuitively not morally responsible. Last, since Pereboom is attempting to show both that the compatibilist conditions for free will are insufficient and that determinism is incompatible with free will and moral responsibility, a single feature of these cases – that Plum's actions are causally determined by factors outside his control – needs to explain why it is that individuals intuitively find Plum not morally responsible. If this intuition is the result of any other aspects of the argument, then Pereboom's argument fails because something independent of the features of determinism would best explain why people judge that Plum lacks moral responsibility. Given that correctly explaining intuitions about Plum is

2. Pereboom asserts that in all four cases Plum satisfies the requirements which Hume (1739/1978), Harry Frankfurt (1971), John Fischer and Mark Ravizza (1998), Jay Wallace (1994), and Alfred Mele (1995; 2006) have argued are necessary for an agent to be considered morally responsible.

3. While it may be impossible for both manipulation to occur and for manipulated agents to meet all compatibilist requirements for moral responsibility (Demetriou 2010), for the purposes of this paper, I will assume these features are compatible with one another.

vital for the success of Pereboom's argument, Pereboom (2014) presents the four-case manipulation argument as an argument for the best explanation.[4]

There is reason to believe that readers are easily confused about what determinism entails (Murray and Nahmias 2014), that readers fail to understand manipulated agents as having all of the necessary compatibilist requirements for moral responsibility (Sripada 2011), and that readers don't actually get the intuition that Plum lacks moral responsibility (Feltz 2013). While these are significant problems for Pereboom's argument, I will focus my attention only on the problem that arises from neglecting to respect other factors that may influence intuitions of non-responsibility. I argue Pereboom's presentation of the four-case argument likely leads to certain, largely unconscious, psychological influences driving intuitions that Plum is not morally responsible. Since the effects of these unconscious psychological influences lead to order effects, I argue that order effects can provide a plausible, and likely better, explanation for why readers get the intuition that Plum is not morally responsible. Pereboom's argument is credibly threatened and potentially undermined by neglecting to ascertain the presence and impact of such influences.

It is important to note that I am not offering a hard-line response and arguing that Plum actually *should* be considered morally responsible in all four cases (McKenna 2008), nor am I taking a soft-line response and arguing that there is a relevant dissimilarity between two of the cases which allows us to consider Plum not morally responsible in Case 1 but morally responsible in Case 4 (Demetriou 2011, Waller 2013). Rather, I take a stance similar to Mele (2005) and call into question Pereboom's explanation for why we find it intuitive that Plum is not morally responsible. We must be certain that the cases are presented in such a way that our intuitions actually track features relevant to the debate before arguing about what intuitions are appropriate for each case. [5]

III. Order Effects as an Alternative Explanation

In this section, I argue that order effects serve as a plausible alternative explanation for what likely motivates judgments of Plum's non-responsibility in the four-case

4. It has been argued that the four-case manipulation argument can be best employed without understanding it as an argument for what best explains intuitions (Mele 2005). I will address this objection later in this paper.

5. Kadri Vihvelin has recently made a similar point and argued that using certain intuitions and thought experiments where is not clear what is being described or when we do not all agree about the verdict, such as manipulation cases, is not helpful for advancing the free will debate.

argument. After providing evidence that the order in which Pereboom's four cases are presented affects judgments about whether Plum is morally responsible, I will discuss specific features and psychological mechanisms which likely lead to order effects occurring in the four-case argument.

Alex Weigmann, Yasmina Okan, and Jonas Nagel (2012) demonstrated that the order in which trolley dilemmas are presented significantly influences judgments of moral permissibility.[6] After presenting participants with five variations of the trolley dilemma, which differed only in what the life-saving action was, they found that the order in which the cases were presented drastically influenced responses to each scenario.[7] Weigmann et al. concluded, "judgments would be most likely transferred if the initial rating was strongly negative" (2012, 825). That is, when readers had a strongly negative judgment towards the first case, this judgment was likely to affect judgments of later cases. This highly negative first case resulted in consistently more negative judgments of moral permissibility relative to judgments of these cases presented on their own. Given that readers have strongly negative reactions to Case 1 in Pereboom's four-case manipulation argument (Feltz 2013), I argue it is highly likely that the order in which these cases are presented by Pereboom has an effect on judgments of Plum's level of moral responsibility in later cases much in the same way Wigmann et al. observed order affected judgments about the trolley dilemmas.

While one might assume the experienced agnostic philosopher would not be affected by the order in which cases are presented, Schwitzgebel and Cushman (2012) found that with respect to moral principles, order of presentation influenced the judgments

6. Trolley dilemmas are scenarios where a trolley train is out of control and on track to run over multiple workers. However, someone has the option of choosing to sacrifice the life of one person to save the multitude.

7. The potentially life-saving actions were: pressing a switch that will redirect the train that is out of control to a parallel track where one person will be run over; redirecting an empty train that is on a parallel track onto the main track to stop the train, running over a person that is on the connecting track; redirecting a train with a person inside that is on a parallel track onto the main track to stop the train; pushing a button that will open a trap door that will let a large person on top of a bridge fall and stop the train; push the large person from the bridge to stop the train.

of philosophers *more* than it did non-philosophers![8] Furthermore, this effect persists among philosophers self-reporting familiarity, expertise, stability, and specialization in ethics (Schwitzgebel and Cushman forthcoming). Not only does this finding suggest that philosophers need to take the salience of order effects seriously, it provides reason for philosophers to take these effects more seriously than others. If it turned out that order effects better explain why we find Plum not morally responsible in Case 4, then Pereboom would fail to provide the best explanation for these intuitions and his argument would be unsuccessful.

Agency-Detection Mechanism

A psychological mechanism that likely guides intuitions and contributes to the effect that order has on judgments regarding Pereboom's four cases is an agency-detection mechanism. Scott Atran (2006) argues that human evolution has naturally selected for an innate and overly sensitive mechanism for detecting agents and agential properties. While this mechanism often beneficially and accurately identifies agents, Atran argues that it also causes humans to wrongly attribute agential properties to nearly any complex or uncertain situation or design. For example, Atran believes this overly sensitive mechanism explains why people often see faces in the clouds and are quick to believe in supernatural beings. This mechanism would become active in Case 1 and correctly lead us to attribute agential properties to the causal determinants of Plum's actions (i.e., the neuroscientists). However, an agency-detection mechanism would likely remain active in later cases when Pereboom replaces these agents with the complex structure of causal determinism, which, importantly, contains no agential properties. If this mechanism remained active, then readers would (perhaps unconsciously) attribute agential properties to the causal determinants of Plum's actions in Case 4. Such attributions would, thereby, alter judgments of Case 4 by confusing the reader about the nature of determinism.[9]

8. Pereboom (2014, 81) states, "…the manipulation argument aims to persuade the natural compatibilist and the agnostic their resistance to incompatibilism is best given up." While it is extremely important to properly recognize who Pereboom's intended audience *is,* who Pereboom's audience *ought to be,* and to what degree such an audience actually exists, I do not have room to adequately address these concerns in this paper.

9. In an unpublished manuscript, Neil Levy makes a similar argument, claiming that Pereboom's four-case manipulation argument only succeeds insofar as it activates an agency-detection mechanism which causes the readers to see determinism in agential terms.

If this overly sensitive agency detection mechanism does, in fact, influence intuitions about Case 4, then the order in which Pereboom presents these cases has an effect on judgments of Plum's non-responsibility. Furthermore, this alternative explanation for intuitions would undermine Pereboom's goal of getting readers to properly understand the causal nature of determinism. Since determinism, and therefore Case 4, does not involve agents or agential properties which influence Plum, it would be misguided for intuitions about Case 4 to be influenced by agency. If intuitions about Plum in Case 4 are motivated by an agency-detection mechanism responding to agency in earlier cases, as I argue they are, then these intuitions are unreliable and cannot be used to motivate Pereboom's argument.

Intent

While the mere presence of agents in Case 1 might cause readers to judge Plum not morally responsible in Case 4, the intent of these agents also appears to contribute to the order effects. Phillips and Shaw (forthcoming) investigated how third-party intent (the intent of agents who causally determine how another agent acts but nonetheless are not involved in the action themselves) influences judgments of moral responsibility. First, they found that the presence of third-party intent does reduce judgments of blame.[10] Second, third-party intent *only* influenced judgments when the agent's actions perfectly match with the intended action. Third, their results suggest that intent affects judgments of moral responsibility by altering the reader's causal perception. If Pereboom's four-case argument successfully alters one's causal perception only because third-party intent is present in earlier cases, then judgments of earlier cases are influencing judgments of later cases, and order effects are thereby produced. If intuitions of Plum's non-responsibility are the result of order effects, then we have an alternative explanation for these intuitions that is deeply problematic for Pereboom's argument.

To see why third-party intent altering judgments would be problematic, consider that according to Pereboom, many people don't see determinism as ruling out the possibility of moral responsibility because they misunderstand the true nature of determinism. To remedy these misconceptions, "the manipulation cases are formulated so as to correct for inadequacy in the extent to which we take into account hidden deterministic causes in our intuitions about ordinary cases" (2014, 95). That is, manipulation cases are intended to expose to us the true causal nature of determinism and they attempt to alter how one

10. These findings are consistent with Robyn Waller's (2013) argument that intent is a relevant difference between cases and affects judgments of moral responsibility.

perceives the causal implications of determinism. Phillips and Shaw's research suggests that manipulation cases can succeed in altering one's causal perception *only* when third-party intent is present and matches the action performed. Therefore, according to Phillips and Shaw's assessment, if a change in causal perception occurs, it must be because readers understand there to be third-party intent present which matched the action. While Pereboom is clearly attempting to change the reader's causal perception, it would be mistaken to alter perceptions by getting readers to understand determinism as having any intent (or, for that matter, any other agential properties) since compatibilists and incompatibilists agree this is the wrong way to conceive of determinism. This suggests that Pereboom elicits the desired intuitions by confusing readers about the true nature of determinism.

While the concern outlined above is certainly problematic for Pereboom's argument, it is worth noting that in order for my argument to succeed, intent doesn't necessarily need to confuse readers about the true nature of determinism. Rather, I merely need to demonstrate that the intent, along with other unconscious psychological influences, lead to order effects influencing judgments and that these order effects explain intuitions of non-responsibility better than the mere fact that Plum's actions are causally determined by factors over which he has no control.[11]

Emotional Responses

Another psychological influence that likely motivates order effects in Pereboom's four-case argument is emotional engagement with features present in Case 1. The first case of the four-case argument involves agential intent, an abnormal bodily violation (brain manipulation), and an abnormal social violation (manipulation). Reading vignettes that contain intent, abnormal bodily violations, and abnormal social violations have been shown to elicit emotional responses (Giner-Sorolla 2011; Haidt 2003). Also, engaging emotionally with such vignettes has been shown both to be correlated with particular moral judgments (Greene 2001), as well as to influence moral judgments (Haidt 2003; Guiseppe et al. 2012) even when these emotions are primed non-consciously and

11. In a response to Mele's criticisms, Pereboom (2014, 82) argues even if these intentional agents, "were replaced by force fields or machines that randomly form in space that have the same deterministic effect on Plum as the manipulators do, the intuition that Plum is not morally responsible persists." While I remain skeptical of this claim, it is interesting that Pereboom chooses not to make this replacement and he only mentions such a possibility after priming the reader with cases involving intentional agents.

automatically (Valdesolo and DeSteno 2006).[12] Furthermore, responding emotionally to a vignette has been shown to affect judgments and behavior continually for some time after reading the vignette (Plaisier and Konijn 2013; He et al. 2013).

In light of such evidence, it seems very likely that readers of Pereboom's four-case argument would have a strongly negative emotional response to Case 1 and that this highly negative response would influence judgments regarding Case 4. Insofar as one's emotions are negatively responding to agential intent, body violations, or social violations, and not to the fact that Plum's actions are causally determined by factors he has no control over, emotional engagement serves as a plausible confounding variable for what explains judgments. That is, if our intuitions about Plum are the result of responding to emotional-priming factors that are irrelevant to determinism, then it isn't a feature of determinism that drives moral judgments, as Pereboom argues. Since features of Case 1 are known to elicit emotional reactions, it seems likely that emotional engagement with features present in Case 1 influence judgments of later cases, thus leading to order effects taking place. These order effects, again, serve as an alternative explanation for intuitions of Plum's non-responsibility in Case 4 and thereby threaten the success of Pereboom's four-case manipulation argument.

In summary, given Pereboom's presentation of his four-case manipulation argument, it is likely that features only present in earlier cases (agents, third-party intent, abnormal body and social violations) are initiating certain unconscious psychological mechanisms that drive judgments of Case 4, thus resulting in order effects. There may be additional psychological influences that drive order effects which I have not discussed. For example, intuitions could also be swayed by one's own demands for consistency across cases, readers having intuitions of non-responsibility simply because Pereboom makes suggestions about what intuitions readers ought to have, or readers agreeing with Pereboom because he is understood to be some kind of authority figure on what one ought to think about these cases. If *any* such influences, either collectively or on their own, better explain why we (or "agnostic" readers) find Plum intuitively not morally responsible, then Pereboom's argument is unsuccessful. Therefore, Pereboom, like anyone else attempting to make claims about what drives intuitions, needs to take unconscious psychological influences seriously. As I have now demonstrated, neglecting

12. Haidt (2001) argues that in most circumstances, emotional engagement is the primary cause of moral reasoning. While this may or may not be the case, for my argument to work, it only needs to be the case that emotional engagement influences judgments of Pereboom's four cases.

to acknowledge seemingly irrelevant influences, such as order effects, can undermine one's entire argument.

IV. Objections

Thus far, I have argued that by failing to recognize salient and largely unconscious psychological influences that have been shown to affect intuitions, Pereboom's four-case manipulation argument likely does not elicit judgments about moral responsibility in a way that is required to support the argument. More specifically, I have argued that the intuition that Plum is not morally responsible is not likely best explained by the fact that Plum's actions are causally determined by factors outside of his control. Rather, these intuitions are more plausibly explained by the presence of order effects that are driven by certain psychological influences which readers are largely unaware of, such as an agency-detection mechanism, third-party intent, and highly negative emotional engagement. I will now entertain objections to my argument.

Order Effects Are Intended

First, one might be tempted to object to my argument by saying something like the following: "Of course order effects sway intuitions in Pereboom's favor. The whole point of the four-case argument is to lead people to understand that the factors that undermine moral responsibility in Case 1 undermine responsibility in Case 4 as well. Therefore, the emotional responses and initial judgments about Case 1 *should* transfer over and influence intuitions about Case 4 so that we think of these cases in the same way and with the same types of attitudes."

In response to this objection, I would first point out that insofar as Pereboom's four-case manipulation argument is to be understood as an argument to the best explanation, the argument only works if Pereboom's explanation is actually the best. Therefore, if the fact that Plum's actions are being determined by factors outside his control is *not* what drives intuitions, then the argument simply doesn't work. Mele (2005; 2008) has argued that readers would judge Plum not morally responsible even if the causation in these cases was indeterministic, and this would show that determinism is not what motivates intuitions about the four cases. If Mele is right and deterministic causation isn't what drives intuitions, then these judgments must be sensitive to other factors within these cases. I presented a few likely candidates for which features of these cases influence intuitions regarding Case 1: the presence of agents, third-party intent, and emotionally responding to manipulation. Furthermore, I provided reason to believe

that if the factors I discuss are what motivate intuitions about Case 1, then it's highly likely that order effects will take place as a result and intuitions of non-responsibility will remain consistent across cases. Therefore, order effects driven by psychological influences that attend to features present in Case 1 serve as a confounding variable for the success of Pereboom's argument if these order effects better account for what motivates the intuition that Plum is not morally responsible.

As a second response to this objection, I'd point out that if order effects are supposed to take place and we are supposed to understand Case 1 and Case 4 in roughly the same way, then Pereboom is likely confusing the reader about the true causal nature of determinism. As discussed earlier, if the intuition that Plum is not morally responsible in Case 4 is residually influenced by the presence of agents or third-party intent in Case 1, then the intuitions about Case 4 are misguided since determinism has no agential properties or intentions.

If it turns out that intuitions about Case 1 are solely, or at least primarily, motivated by the fact that Plum's actions are causally determined by factors outside his control, and if after reading the four-case argument readers are not at all confused about determinism, then judgments regarding Case 4 being influenced by order effects would not be problematic. However, as I have now argued, it seems extremely unlikely that judgments are best explained by the single feature Pereboom addresses, given the many other features present in Case 1 that are known to engage psychological mechanisms that lead to order effects and alter judgments of later cases. Furthermore, it seems plausible that readers are conflating features such as agency and intent with determinism in Case 4, thus confusing the reader about the true nature of determinism. Work in experimental philosophy has provided evidence of such confusion (Murray and Nahmias 2014; Sripada 2011).

Explaining Intuitions Is Unimportant

A second objection to my argument is that by presenting Pereboom's four-case argument as an argument to the best explanation, I am misrepresenting it. Thus far, I have been assuming that Pereboom's explanation for intuitions about the four cases is a central feature of his argument. Nonetheless, it's possible that one can conclude Plum is not morally responsible without offering any explanation at all for what drives these intuitions. In response to this objection, I argue that this alternative understanding of the four-case manipulation argument, besides running counter to Pereboom's stated intentions, is extremely problematic.

In Pereboom's most recent presentation of his four-case argument, he argues,

> It's highly intuitive that Plum is not morally responsible in Case 1, and there are no differences between Cases 1 and 2, 2 and 3, and 3 and 4 that can explain in a principled way why he would not be responsible in the former of each pair but would be in the latter. We are thus driven to the conclusion that he is not responsible in Case 4. The salient factor that can plausibly explain why Plum is not responsible in all of the cases is that in each he is causally determined by factors beyond his control to decide as he does. This is therefore a sufficient, and I think also the best, explanation for his non-responsibility in all of the cases. (2014, 79)

This passage might lead one to assume Pereboom's argument is similar to other manipulation arguments, which can very roughly be formulated in the manner below. I will refer to this formulation as MA.

(P1) Plum is not morally responsible in Case 1.

(P2) There are no differences between cases that are relevant to moral responsibility.

(C) Therefore, Plum is not morally responsible in Case 4, and since Plum in Case 4 is no different from any agent in a deterministic universe, no agents in a deterministic universe are morally responsible either.

MA seems to get the conclusion Pereboom desires without employing any premises that explain intuitions. While one *could* present Pereboom's argument in a way that does not make use of his explanation for intuitions, I would argue that this understanding of Pereboom's argument would be problematic.

Though there may be other problems with this kind of formulation, I will focus my attention on the fact that it draws a conclusion about moral responsibility directly from an intuition about moral responsibility: Plum *is intuitively* not morally responsible in Case 1. Therefore, Plum *is* not morally responsible in Case 1. This reasoning is what motivates P1 of MA. Nonetheless, if this move is permitted then compatibilists could simply employ the same reasoning and argue that because they find persons in deterministic universes *intuitively* morally responsible, then these agents must actually be morally responsible

(King 2013). Furthermore, if such reasoning is permitted, then debates about free will and moral responsibility would be reduced to a battle of intuitions instead of being won via philosophical argumentation. While this reduction is undesirable and would likely be unfruitful, one might argue this is what Pereboom has in mind. For instance, in his response to McKenna's criticisms of the four-case argument, Pereboom (2005, 242) suggests we "let the intuitions fall where they may."

Appealing to intuitions without any explanation for what drives these intuitions may be useful if virtually all readers have the same intuition about the cases presented. However, this universality doesn't seem to occur with Pereboom's four-case manipulation argument (Feltz 2013) or similar cases involving manipulation or determinism (Murray and Nahmias 2014; Nichols and Knobe 2007; Sripada 2011). Given that intuitions about these cases are not uniform, the only ways to avoid a stalemate is to explain what drives intuitions about P1 or simply provide a separate, substantive philosophical argument which justifies P1.

I assume that Pereboom intends to avoid such a stalemate and the related methodological issues which arise from understanding his argument to be formulated similar to MA. There is good reason to consider Pereboom's explanation for what drives intuitions as a significant aspect of his four-case argument, since Pereboom himself explains this is how the argument ought to be understood in a footnote. He states,

> Al Mele (2006) argues that a manipulation argument against compatibilism need not be cast as an argument to the best explanation. I doubt that this is so. True, the argument can be represented without a best-explanation premise, but such a representation will not reveal its real structure. By analogy, the teleological argument for God existence can be represented as a deductive argument, but its real structure is an argument to the best explanation for biological order in the universe. The fact that the real structure of a manipulation argument against compatibilism is an argument to the best explanation becomes clear when one considers compatibilist objections to it—that, for, example, the non-responsibility intuitions can be accounted for by manipulation of a certain sort and not by causal determination. (2015, 79-80)

Here Pereboom makes it clear that his argument is one in which the explanation of intuitions is paramount. Furthermore, he states that the way one should object to his argument is by providing an alternative explanation for what causes intuitions of Plum's non-responsibility. This is exactly what I have attempted to do in this paper.

As a final note, I'd point out that the claims Pereboom and myself make about what best explains intuitions are empirical claims. It's possible to manipulate the features of these cases and determine what does and does not motivate intuitions. Furthermore, we can test whether, after reading Pereboom's four-case argument, readers correctly understand the true causal nature of determinism. If it turns out intuitions of Plum's non-responsibility are directly driven by the fact that Plum's actions are causally determined by factors outside his control and, if after reading all four cases, readers understand exactly what determinism entails, then Pereboom's argument would successfully avoid my criticisms in this paper. I doubt, however, that this is what we would find and hope to investigate these matters empirically in the future.

V. Conclusion

The goal of this paper was to demonstrate that arguments which appeal to intuitions about thought experiments and hypothetical cases must acknowledge the many psychological influences that subtly motivate intuitions. I argued that influences, such as order effects, can affect judgments to the extent that arguments which employ these cases are unsuccessful. Without ensuring that our intuitions are tracking relevant features of an argument, intuitions regarding thought experiments will likely be unreliable and, therefore, fruitless for the purposes of philosophical discussion. To exemplify these concerns, I presented Derk Pereboom's four-case manipulation argument. I have provided evidence that suggests intuitions about these four cases can better be explained by order effects than by recognizing that Plum's actions are causally determined by factors outside of his control. Since it may be the case that what best explains intuitions of Plum's non-responsibility across all four cases is not that Plum's actions are causally determined by factors outside his control, order effects serve as a plausible alternative explanation for what drives intuitions. If what drives intuitions about Pereboom's hypothetical cases are factors irrelevant to causal determinism, as I argue is the case, then by failing to correctly identify what motivates intuitions about his four cases, Pereboom's argument is unsuccessful.

My suggestion to consider alternative psychological explanations, such as order effects, when explaining what motivates intuitions does not solve the potential problem of unreliability that arises as a result of intuitions differing across demographics. However, I have provided evidence that intuitional unreliability, in the sense that intuitions are sensitive to trivial features of hypothetical cases and thought experiments, is problematic when one's explanation for what motivates these intuitions is incorrect. One must take

seriously the fact that intuitions are influenced by many seemingly irrelevant factors when attempting to use thought experiments or hypothetical cases to provide support for an argument. Just as a good scientist considers all confounding variables before claiming to know the cause of a certain event, philosophers must address potential confounding factors for intuitions.

References

Alexander, Joshua, Ronald Mallon, and Jonathan Weinberg. 2014. "Accentuate the Negative." In *Experimental Philosophy* Volume 2, Edited by Joshua Knobe and Shaun Nichols, 31–50. New York: Oxford University Press.

Atran, Scott. 2006. "Religion's Innate Origins and Evolutionary Background." In *The Innate Mind: Culture and Cognition*, edited by Peter Carruthers, Stephen Laurence, and Stephen Stich, 302–317. Oxford: Oxford University Press.

Buckwalter, Wesley, and Stephen Stich. 2011. "Gender and the Philosophy Club." *The Philosophers' Magazine* 52: 60–65.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: The MIT Press.

Demetriou, Kristin. 2010. "The Soft-Line Solution to Pereboom's Four-Case Argument." *Australasian Journal of Philosophy* 88 (4): 595–617.

Feltz, Adam. 2013. "Pereboom and premises: Asking the right questions in the experimental philosophy of free will." *Consciousness and Cognition* 22 (1): 53-63.

Fischer, John Martin, and Mark Ravizza 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20.

Giner-Sorolla, Roger, and Pascale Sophie Russell. 2011. "Moral anger, but not moral disgust, responds to intentionality." *Emotion* 11 (2): 233–240.

Greene, Joshua D. 2011. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–2108.

Guiseppe, Ugazio, Claus Lamm, and Tania Singer. 2012. "The role of emotions for moral judgments depends on the type of emotion and moral scenario." *Emotion* 12 (3): 579–590.

Haidt, Jonathan. 2001. "The Emotional Dog And Its Rational Tail: A Social Intuitionist Approach To Moral Judgment." *Psychological Review* 108 (4): 814–834.

Haidt, Jonathan. 2003. "The moral emotions." In *Handbook of Affective Sciences*, edited by Richard J Davidson, Klaus R Sherer, and H. Hill Goldsmith 852–870. Oxford: Oxford University Press.

Haji, Ishtiyaque. 1998. *Moral Accountability*. New York: Oxford University Press.

Haji, Ishtiyaque. 2009. *Incompatibilism's Allure: Principal Arguments for Incompatibilism*. Peterborough ON: Broadview Press.

He, J.; X. Jin, M. Zhang, X. Huang, R. Shui, and M. Shen. 2013. "Anger and selective attention to reward and punish children." *Journal of Experimental Child Psychology* 115 (3): 389-404.

Hume, David. (1739) 1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

King, Laura A., and Joshua A Hicks. 2011. "Subliminal mere exposure and explicit and implicit positive affective responses." *Cognition & Emotion* 25 (4): 726–729.

King, Matt. 2013. "The Problem With Manipulation." *Ethics* 124 (1): 65–83.

Levy, Neil. Unpublished manuscript. "Manipulating the Reader: Manipulation Arguments and Agency Detection."

Li, Wen, Richard E. Zinbarg, Stephan G. Boehm, and Ken A. Paller. 2008. "Neural and Behavioral Evidence for Affective Priming from Unconsciously Perceived Emotional Facial Expressions and the Influence of Trait Anxiety." *Journal of Cognitive Neuroscience* 20 (1): 95–107.

McKenna, Michael. 2008. "A hard-line reply to Pereboom's four-case manipulation argument." *Philosophy and Phenomenological Research* 77 (1): 142–159.

Mele, Alfred. 1995. *Autonomous Agents*. New York: Oxford University Press.

Mele, Alfred. 2005. "A critique of Pereboom's 'four-case argument' for incompatibilism." *Analysis* 65 (1): 75-80.

Mele, Alfred. 2006. *Free Will and Luck*. New York: Oxford University Press.

Mele, Alfred. 2008. "Manipulation, Compatibilism, and Moral Responsibility." *The Journal of Ethics* 12 (3–4): 263–286.

Mlodinow, Leonard. 2012. *Subliminal: how your unconscious mind rules your behavior*. New York: Pantheon Books.

Murray, Dylan, and Eddy Nahmias. 2014. "Explaining Away Incompatibilist Intuitions." *Philosophy and Phenomenological Research* 88 (2): 434–467.

Nichols, Shaun, and Joshua Knobe. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *NOUS* 41 (4): 663–685.

Pereboom, Derk. 2005. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy* 29 (1): 228–247.

Pereboom, Derk. 2001. *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2014. *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Petrinovich, Lewis, and Patricia O'Neill. 1996. "Influence of Wording and Framing Effects on Moral Intuitions." *Ethology and Sociobiology* 17 (3): 145–171.

Phillips, Jonathan, and Shaw, Alex. Forthcoming. "Manipulating Morality: Third-Party Intentions Alter Moral Judgments by Changing Causal Reasoning."

Plaisier, Xanthe S., and Konijn, Elly A. 2013. "Rejected by peers—Attracted to antisocial media content: Rejection-based anger impairs moral judgment among adolescents." *Developmental Psychology* 49 (6): 1165-1173.

Schwitzgebel, Eric, and Fiery Cushman. 2012. "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind & Language* 27 (2): 135–153.

Schwitzgebel, Eric, and Fiery Cushman. Forthcoming. "Professional Philosophers' Susceptibility to Order Effects and Framing Effects in Evaluating Moral Dilemmas."

Sinnott-Armstrong, Walter. 2008. "Framing Moral Intuition." In *Moral Psychology, Vol 2. The Cognitive Science of Morality: Intuition and Diversity*, 47–76. Cambridge, MA: MIT Press.

Sripada, Chandra. 2011. "What makes a manipulated agent unfree?" *Philosophy and Phenomenological Research* 85 (3): 1–31.

Tobia, Kevin P., Gretchen B. Chapman, and Stephen Stich. 2013. "Cleanliness is Next to Morality, Even for Philosophers." *Journal of Consciousness Studies* 20 (11–12): 195–204.

Todd, Patrick. 2012. "Defending (a modified version of) the Zygote Argument." *Philosophical Studies* 164 (1): 189–203.

Valdesolo, Piercarlo, and David DeSteno. 2006. "Manipulations Of Emotional Context Shape Moral Judgment." *Psychological Science* 17 (6): 476–477.

Vaesen, Krist, Martin Peterson, and Bart Van Bezooijen. 2013. "The Reliability of Armchair Intuitions." *Metaphilosophy* 44 (5): 559–578.

Vihvelin, Kadri. "How Not to Think about Free Will." *Journal of Cognition and Neuroethics* 3 (1).

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Waller, Robyn. 2013. "The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument." *Philosophia* 42 (1): 209–222.

Weigmann, Alex, Yasmina Okan, and Jonas Nagel. 2012. "Order effects in moral judgment." *Philosophical Psychology* 25 6: 813–836.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. 2001. "Normativity and epistemic intuitions." *Philosophical Topics* 29 (1-2): 429–460.

Weinberg, Jonathan, Stacey Swain, and Joshua Alexander. 2008. "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp." *Philosophy and Phenomenological Research* 77 (1): 138–55.

Journal of Cognition and Neuroethics

Identity and Freedom

A.P. Taylor
North Dakota State University

David B. Hershenov
University at Buffalo

Biographies

David B. Hershenov is a professor and chair of the philosophy department at the University at Buffalo. He works primarily on issues at the intersection of personal identity and bioethics.

A.P. Taylor is a lecturer in philosophy at North Dakota State university. He works on personal identity and the metaphysical foundations of well-being.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Taylor, A.P., and David B. Hershenov. 2015. "Identity and Freedom." *Journal of Cognition and Neuroethics* 3 (1): 383–391.

Identity and Freedom

A.P. Taylor and David B. Hershenov

Abstract
Few philosophers will think someone is free and responsible if he forms his intentions while either thinking that he is someone else, say Barack Obama, or is considering only Obama's values and interests while kept ignorant of his own. We'll argue that there is an analogous problem for all the major materialist theories that understand human persons to be essentially thinking beings that physically overlap but are distinct from human animals. The source of the problem is that these accounts posit that there exist more than one entity possessing the same brain – the person and the animal. Since the former can use the brain to think, so can the latter. If there's more than one overlapping thinker and their interests diverge then there'll arise the problem that each cannot exercise his free will without undermining the other's free will. We conclude that, among materialist theories of personal identity, only an animalist metaphysics that identifies material persons and animals will provide a necessary condition for our being free.

Keywords
Free Will, Animalism, Personal Identity, Materialism

Introduction

It would certainly be unwelcome if one's preferred metaphysics of the person often makes free will and moral responsibility impossible. We contend that this is the fate of all the major materialist theories that understand human persons to be essentially thinking beings that physically overlap human animals. The source of the problem is that these accounts posit that there exist more than one entity possessing the same brain—the person and the animal. If the former can use the brain to think, so can the other. As a result, such theories are afflicted by *The Problem of Too Many Thinkers*.

Our focus is on an overlooked moral version of the Problem of Too Many Thinkers. If there is more than one overlapping thinker then there'll arise the problem that each cannot freely and responsibly act in a way that respects the exercise of the other's freedom. The trouble comes about because the overlapping thinkers can't simultaneously think and act on their interests and govern their lives in accordance with their values. And their interests and values will diverge due to their having different persistence conditions. We will show that there could be many occasions where only the person will give her free consent. The human animal overlapping the person won't freely consent to the same

intention or action because he will either wrongly think that he is the person or will just be considering what is in the person's interests.

Our contention is that the only prominent materialist account to avoid such problems is the animalist theory that identifies the human person and the human animal and adopts a sparse ontology. We reach this conclusion in part because we accept a methodology in which ethical considerations and action-theoretic claims can weigh against certain metaphysical accounts of the person. So we aren't forced by methodological principles to claim that certain metaphysical conceptions of the person show that there is no free and responsible action; rather, we can plausibly claim that practical considerations strongly suggest that those metaphysical approaches to personal identity are false.

One reason why we believe practical considerations should be included in the weighing of reasons in favor and against a metaphysical theory is that if there are moral truths then they must be consistent with metaphysical truths. If one adopts a metaphysic in which thinking beings overlap then one must reject certain seemingly obvious moral truths like we ought to respect the free choice and bodily integrity of beings like ourselves. We find it plausible that such considerations should tilt the scales against that metaphysics rather than show such core moral principles to be false. But even if one is an anti-realist about ethics and thus under no pressure to make metaphysical truths cohere with moral truths for there are not any of the latter, there are still true action-theoretic claims about free action that we think a metaphysic should accommodate. For instance, if a metaphysic makes it impossible for everyone intelligent enough to understand these sentences to also freely endorse and act upon their interests and values, then that too provides a reason to doubt the metaphysics rather than accept the impossibility of such creatures being free.

The Moral Problem of Too Many Thinkers

We believe that greater success in resolving *The Problem of Too Many Thinkers* is the closest there is to a criterion to choose between competing metaphysical theories of the person. If a theory implies that there exists more than one thinker under your clothes, then that is a major strike against the theory. This reason may not be strong enough on its own to warrant rejecting the theory outright, but it will greatly weigh against it, tilting the scales further in the direction of a view that avoids the problem.

Let's assume that persons are essentially thinking beings that are spatially coincident but numerically distinct from animals. The problem which arises is that if the person can

think, then it would seem that the animal should also be able to think since it shares the same brain. Olson (1997; 2002; 2007) highlights the epistemic problem that arises if both the person and the animal can think, then you have no reason to think you are the person rather than the animal. Either you or your co-located thinker will be wrong when you both say "I am essentially a person." The first person pronoun refers to each of the speakers and so one is falsely predicating personhood to itself. How can you be so sure that you are not the deluded animal making the erroneous self-ascription?

Noonan (2010) attempts to avoid the epistemic problem by endorsing what has become known as *Pronoun Revisionism*. Noonan suggests that to have thoughts about one's thoughts is not enough to make an entity a person, rather an individual must have the appropriate psychological persistence conditions. That is, the person goes out of existence if he loses certain psychological capabilities. So the thinking animal is never a referent of the personal pronoun "I" for the term doesn't pick out any entity thinking or uttering the word "I", rather it just refers to the person.

However, even if Noonan is right about the animal's use of the personal pronoun, this will not be enough to mitigate the ethical problems. While we'll draw mostly upon bioethical examples involving what is known as informed consent, readers can easily imagine similar scenarios in other domains that would likewise undermine free will. If there are non-persons such as human animals that can't refer to themselves with the first-person pronoun, then how can they be said to freely agree to any immediate treatment or make provisions for their future with say a living will? While we don't have a favored theory of free will to expound, it would seem safe to say that one couldn't be free if unable to reflect upon one's interests, desires, values, intentions as *one's* own, and then act on the basis of the reasons they provide. If we assume pronoun revisionism, a problem is that the animal is thinking about the person's interests as the person does, for the animal refers to the person when it uses first person pronouns to entertain thoughts of the following types: "*I* would prefer such and such a course of treatment for it increases *my* well-being" or "I would prefer to forgo all treatments so I can lead the remainder of my life in accordance with my values." Our worry is that the animal might have interests and values that are not the same as those of the person because of their different persistent conditions. Thus the animal's choosing to act on the basis of what the person has reason to do cannot be understood to be a free and responsible action of the animal who didn't reflect upon the action qua animal, i.e., did not think of himself as an animal engaged in that action. A similar lesson can be drawn in the absence of pronoun revisionism if the animal uses the first person pronoun to self-refer but is ignorant of his kind membership, wrongly thinking he is the person. He will be choosing actions on the

basis of the person's interests and values and so the action cannot be said to emerge from him in a way characteristic of freedom.

We'll provide examples below in which overlapping thinkers won't be self-governing on any of the leading theories of free will. On psychological accounts like those of Noonan, Shoemaker, Parfit, Baker and Hudson, persons go out of existence with the loss of certain sophisticated psychological capacities while human animals can survive the loss of such capacities, existing in impaired mental states. We contend that the animal might have an interest in continuing to exist in a childlike state after the rational, self-conscious person has ceased to exist as a result of injury or stroke. Conversely, persons might have interests that are not strictly those of our animals. A conflict can be generated if there is an experimental drug that may prevent the further decline into Alzheimer's disease, but will far more likely kill its user. The person, who inevitably goes out of existence with the loss of self-consciousness, might think she has nothing to lose since either the disease or the drug's unwanted side effect will end her existence. However, it may be in the interest of the human animal not to take the drug since it could survive with the minimal sentience of late stage Alzheimer's disease.

We don't think such a choice could be considered free for *both* the animal and the person on any of the leading accounts of freedom. It doesn't matter if such accounts stress the endorsement of desires that we act on by higher order desires or values (Dworkin 1970; Frankfurt 1971; Watson 1975), emphasize the history of how those higher order attitudes arose (Wolf 1990), insist upon choices meshing with long term plans (Bratman 2010), require a reason responsiveness and a mechanism that is sensitive to reasons (Fischer and Ravizza 1998), or insist that the agent exercise his causal power to choose between alternatives regardless of antecedent circumstances (Clark 1996; O'Connor 2008). The overlapping thinkers in the above and below scenarios could consider in succession that they were the person and then the animal, the result being that if they first each thought they were an animal they would endorse different acts, be alienated from different parts of a shared history, have divergent long term plans, be sensitive to different reasons, and exercise agent causation differently from how they would if they thought they were the person.

The Alzheimer's drug isn't the only scenario where free will cannot be exercised by overlapping thinkers. Conflicts between the person and the animal could prevent them from both *freely* endorsing the same advanced directive. For instance, the person may not want his resources to be spent on sustaining an organism with dementia with whom it is not identical. That person would have written a very different advanced directive if he had thought he was the animal. Or the person might leave directions to try an extremely

dangerous experimental treatment if his Alzheimer's Disease progresses to a certain state. But the treatment would be contrary to the animal's interests. So the advanced directive written by the animal and the person while the animal thought it was the person or only considered the latter's interests will not do justice to the animal's freedom.

We can generate other infringements due to the animal and person's different interests due to their different persistence conditions. Assume the person and the animal both support donating organs at their deaths but not before. Let's add that they even believe they are morally obligated to engage in *directed donation* and bestow organs upon an ailing friend or relative after they die. However, the possibility of the animal and person's deaths occurring at different times could prevent the full realization of their seemingly shared value. The person is essentially a thinking being and the animal is essentially a living being and so the criteria for their deaths would diverge. The problem is that when the animal dies after its respiration and circulation have irreversibly ceased, less of its organs may be viable for transplant than if organs were taken when just consciousness was lost irreversibly with the onset of a persistent vegetative state. And in the case of directed donation, the person might not be able to donate upon her death because the organism is still alive and doesn't pass away until long after the needy friend or relative dies.

It is important to realize that these types of conflicts aren't the standard conflicts of interests between free parties where say a government health official doesn't allow the person to have the experimental drug that he covets, or a court rules that a health insurance company needn't provide payment for the expensive treatment that the patient is interested in, or a doctor refuses to undertake the risky procedure that the patient wants. Each of these individuals can freely formulate an intention and act upon it even if someone else later prevents their action from producing the desired results: the acquisition of the drug requested, the petitioned for payment, the provision of the sought after procedure. Rather, it is impossible for the overlapping animal and person to *simultaneously* freely endorse an intention that *both* then act upon. Nor do they each have free control over a personal realm, their body. While you can choose to take a risky experimental drug that I can refuse to take, the spatially located person and animal cannot each make and act on their own choice. If one takes the drug, the other does so as well. If the person donates multiple organs upon his death with the loss of the appropriate mental capacities, the animal will be killed when his vital organs are taken. Conflicts like these make it impossible to respect the bodily integrity of both. So we are not presenting just another instance of the typical problem where someone's freely

endorsed intentions and acts are foiled by the freely produced preferences and deeds of others in the society—and without any rights being violated.

We don't think one can escape this by appealing to Parfit's famous claim that identity isn't what matters, our prudential-like concern being only with psychological connections to a future person regardless of whether we are one and the same person or distinct persons.[1] If Parfit were right, the overlapping human animal and person would have the same interests and thus would not choose differently. Parfit bases his claim on cases like those involving Adam's cerebrum fissioning and both cerebral hemispheres transplanted into different bodies B and C. Adam would have survived if just one of the hemispheres was successfully transplanted, the other destroyed upon removal. So Parfit reasons that having these two equally good psychological successors is as good as ordinary survival (no fission and no transplants). The claim that identity doesn't matter depends upon Parfit holding an account of identity involving a uniqueness clause. Parfit's criterion for personal identity across time is that it consists of i) the appropriate psychological relation R and ii) being uniquely the possessor of such relations. Since this *uniqueness clause*, aka no *branching clause*, is trivial and satisfied by what is extrinsic to us, Parfit insists it can't be what matters to us. So it must be the other component of the personal identity criterion, the psychological relation R, that matters to us. This led Parfit to his famous conclusion that identity doesn't matter.

One reason for our skepticism about Parfit's thesis is that it doesn't mesh with our reactions to torture and death following a great change in our psychology due to a stroke. It doesn't seem that we now will view the later torment and death as being less bad since we are less psychologically connected to the being after the stroke than we would be if there had been no damaging stroke.

Secondly, Parfit's claim that identity doesn't matter depends upon his interpretation of a fission scenario that violates the rationale behind *the only x and y rule*. That rule does not allow that our identity in the future can be determined by whether there are two or more equally good candidates as there would be in cases of fission. The rule restricts questions of whether x is identical to y to the internal relationship between x and y, the existence of a z being irrelevant. The rationale for the rule is that there should not be unexplained existences where entities owe their existence to other beings despite the absence of a causal connection between them (Hawley, 2005). The problem with Parfit's cerebral fission and transplant case is that the person in body B would not be there if it wasn't for the existence of the person in body C likewise being psychologically

1. See Parfit (1984).

continuous with Adam. So the person in Body B owes his existence to the person in body C, and vice versa, but there are no causal connections between person in body B and the person in body C despite the existence of each playing a role in the creation or sustaining of the other. So if unexplained existences are to be avoided, then the criterion for identity should not involve a uniqueness rule and psychological relation R. But it is only the extrinsic and trivial features of the uniqueness clause that leads Parfit to the conclusion that only psychological relations matter. If he is not allowed to introduce the uniqueness rule into the account of identity, then fission can't show that identity doesn't matter, merely psychological relation R is of importance. So Parfit's thesis can't save autonomy in the above cases by giving the overlapping thinkers the same interests.

Conclusion

Thus if you're a materialist and care about freedom and responsibility, then you'd better identify yourself with your animal. So this gives us an additional and rather weighty reason to put on the metaphysical scale, perhaps tilting it in favor of the view that we are animals. The animal is the person. And there aren't any other thinkers overlapping the animal.

If you don't believe that moral or action-theoretic considerations should be given any weight when considering rival metaphysics, then we suggest that you come up with a radically new ethics. It will be an ethics that downplays satisfying free choice, and autonomous control over one's body due to the recognition of the divergent interests and values of overlapping entities. The new ethics will recommend some sort of compromise between the interests and values of those individuals now being counted by the latest metaphysical census, no doubt abandoning many of our currently established rights in the process. But that is the topic for another paper, or rather book, and one that we hope no one ever has to write.

References

Bratman, Michael. 2005. "Planning Agency, Autonomous Agency." In *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, 33–57. Cambridge: Cambridge University Press.

Clark, Randolph. 1996. "Agent Causation and Event-Causation in the Production of Free Action." *Philosophical Topics* 24 (2): 19–48.

Dworkin, Gerald. 1970. "Acting Freely." *Nous* 4 (November): 367–383.

Fischer, John, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge: Cambridge University Press.

Frankfurt. Harry. 1971. "Freedom of the Will and the Concept of the Person." *Journal of Philosophy* 68 (1): 5–20.

Hawley, Katherine. 2005. "Fission, Fusion and Intrinsic Facts." *Philosophy and Phenomenological Research* 71 (3): 602–621.

Noonan, Harold. 2010. "Persons, Animals and Human Beings." In *Time and Identity*, Edited by Joseph Keim Campbell, Michael O'Rourke and Harry S. Silverstein, 185–208. Cambridge: MIT Press.

O'Connor, Timothy. 2008. "Agent-Causal Power." In *Dispositions and Causes,* edited by Toby Handfield, 366–388. Oxford: Oxford University Press.

Olson, Eric. 1997. *The Human Animal: Identity without Psychology*. Oxford: Oxford University Press.

Olson, Eric. 2002. "Thinking Animals and the Reference of 'I.'" *Philosophical Topics* 30 (1): 189–207.

Olson, Eric. 2007. *What Are We? A Study in Personal Ontology.* Oxford: Oxford University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford Clarendon Press.

Watson, Gary. 1975. "Free Agency." *Journal of Philosophy* 72 (April): 205–220.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Journal of Cognition and Neuroethics

How Not To Think about Free Will

Kadri Vihvelin
University of Southern California

Biography

Kadri Vihvelin is Professor of Philosophy at the University of Southern California. Her research focuses on topics in metaphysics and the philosophy of mind and action, including causation, counterfactuals, dispositions, free will, time travel, and mental causation. She is the author of *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, Oxford University Press, 2013.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2015. Volume 3, Issue 1.

Citation

Vihvelin, Kadri. 2015. "How Not To Think about Free Will." *Journal of Cognition and Neuroethics* 3 (1): 393–403.

How Not To Think about Free Will

Kadri Vihvelin

Abstract

Our belief that we have free will is one of those entrenched beliefs of commonsense that no one denies until they start doing some philosophy. Some say that our belief that we have free will is incompatible with the existence of truths about our future actions. That's a mistake—and is generally agreed (by philosophers, at least) to be a mistake—the fatalist's mistake. Others say that free will is incompatible with determinism. They say that if determinism turned out to be true, our common sense belief would turn out to be false. Because our common sense belief is so firmly entrenched, some think that we are entitled to conclude that determinism must be false, and must be false in quite specific ways to give us the "elbow room" or the "leeway" or the "robust alternative possibilities" needed for free will. Others think we have no right to reason, from the basis of a commonsense belief to the falsity of something that science tells us (or at least might tell us). But they also assume that the truth of determinism is incompatible with the truth of our common sense belief that we have free will. But is our common sense free will belief really incompatible with determinism? This I take to be the problem of free will and determinism. It's a problem that's been around for quite awhile. I claim, in my book (Causes, Laws, and Free Will: Why Determinism Doesn't Matter, OUP 2013) to have solved it. I'm not going to talk about my solution. What I want to talk about is something that comes under the heading of methodology. I want to talk about how we should think (and talk and write) about free will. But I will begin, as my title suggests, on a more negative note. I offer the following list of don't's for the free will philosopher: don't change the subject, don't do thought experiments, don't rely on intuitions, don't confuse "that's really strange" with "that's impossible," don't worry about hard cases, and don't analyse.

Keywords

Free will, determinism, fatalism, intuitions, thought experiments, Frankfurt examples, Manipulation argument

How Not To Think about Free Will

We've got free will. I'm able to raise my arm—I just did. Now I'm not doing it. But I'm still able to do it. And it isn't just true that I'm able to raise my arm even when I'm not raising it; it's also true that I'm able to choose to raise my arm even when I'm not choosing to do it. And the same goes for lots of other things that we don't do but can do. We are able to do much more than we actually do. We have unexercised abilities, unexercised powers of causing—call them powers of agent-causation, if you want to give them a fancy name. We are able to make choices, on the basis of reasons and reasoning, whether or not we actually do so. We are able to try to do lots of things, whether or not we actually do try. We are able to acquire new beliefs, even if we are too lazy to do the reading or thinking to do so. We are surrounded by unactualized possibilities; we have

abilities we don't exercise (perhaps some abilities we never exercise). We don't have to do what we do. We are able to do otherwise.

I take this to be a fact. Call it 'the Free Will fact.' No one denies it. Well, not quite. It's one of those facts of commonsense that no one denies until they start doing some philosophy.

Some say that the Free Will fact isn't compatible with the existence of truths about the future. They say that if it was "already true" last week—or last month, or last year, or ten billion years ago—that I would fly to Flint and give this talk today, then that's something I had to do. I never had a choice; I was never able to do otherwise. That's a mistake—and is generally agreed (by philosophers, at least) to be a mistake—the fatalist's mistake.

Others say that the Free Will fact is not compatible with determinism. They say that if determinism turned out to be true, this ordinary fact would not be a fact. Some say that we are entitled to conclude that determinism *must be false*, and must be false in quite specific ways to give us the "elbow room" or the "leeway" or the "robust alternative possibilities" needed for free will. Others think we have no right to reason from a commonsense belief to the falsity of something that science tells us (or at least might tell us). But they also assume that the truth of determinism is incompatible with the Free Will fact.

But is the Free Will fact *really* incompatible with determinism? This I take to be the problem of free will and determinism. It's a problem that's been around for quite awhile. I claim, in my book (Vihvelin 2013) to have solved it. (In case it isn't obvious, I am a compatibilist.) I'm not going to talk about that today.

What I want to talk about is something that comes under the heading of methodology—I want to talk about *how* we should think (and talk and write) about free will. But since my time is very short, I will focus on the negative. I will begin by saying how we should *not* think (or talk or write) about free will. Think of these as Rules for the free will philosopher.

First Rule: Stick to the subject—free will.

Don't start talking about something else instead. Don't start talking about moral responsibility. Free will is necessary, but not sufficient, for moral responsibility. Free will is common—wise people have it, foolish people have it, some say that babies and many nonhuman animals have it. Moral responsibility is not so common—no one thinks that babies are morally responsible and there is lots of controversy about when adults are responsible, and more controversy about whether anyone is ever responsible, or whether the concept of moral responsibility even makes sense. But most of this controversy has

nothing to do with free will. We might agree that everyone in this room has free will and would have free will even if determinism turned out to be true. It would not follow that any of us is morally responsible or even that it is possible for anyone to ever be morally responsible. So let's keep this firmly in mind, when we talk about free will, and not slip into those dangerous phrases like "moral freedom," or "the free will that grounds (or justifies or suffices for) moral responsibility."

Or, at least, let's avoid talking in these ways if we hope to make any progress in figuring out what to say about the problem of free will and determinism.

Second Rule: Avoid thought experiments.

Don't get me wrong. Thought experiments are often a useful tool—sometimes a thought experiment is just what's needed to correct a mistake based on failure of imagination.

For instance, if someone says that you must be awake to be morally responsible, then we can show this false by telling a story about a night watchman who falls asleep on the job, so a burglary occurs on his watch. He was asleep when it happened, but he is still responsible because he could and should have been awake. This is a successful thought experiment but note that it has two ingredients—we can all understand what is being described and we all agree about the verdict. It works because it spells out a possibility we had not thought of, or had forgotten about. (It's a counterexample.)

A more complicated example of a good thought experiment is the story that Sydney Shoemaker told to refute the claim that there can be no time without change (Shoemaker 1969). Shoemaker told a story about a possible world in which there are three distinct regions, each of which experiences a local freeze (a yearlong period when there is no change) at regular intervals. And because the freezes happen in a regular pattern for the entire history of that universe, there is good inductive evidence that every sixty years there will be a global freeze. This is a good thought experiment because it is perfectly clear what is being described, and because the story makes us aware of a possibility that, until Shoemaker described it, had not occurred to us.

Unfortunately the free will literature is filled with example of bad thought experiments.

Manipulation Arguments are bad thought experiments. These are stories in which we are invited, say, to imagine people who are "just like us" except that everything they think and do is "remote controlled" by evil neuroscientists. They come in different varieties,[1] but

1. For one well-known example, see Derk Pereboom's description of Professor Plum in Pereboom 2001, 112–113.

they all suffer from the problem of under-description. It is not at all clear what we are being asked to imagine. And they suffer from the problem that people disagree about the verdict—they haven't changed the minds of any compatibilists. So what exactly is the point? Perhaps to make vivid to the uninitiated what a deterministic universe is like? But we don't need a story about manipulation to do that. Whenever I teach my semester long course on Free Will and Determinism, I succeed in depressing my students for several weeks when I tell them what the thesis of determinism is and gradually convince them that they can't just dismiss it, that its truth—or something close enough—is a live possibility. Why are they depressed? Because they think that determinism rules out free will. That, I believe, is a mistake; others disagree. But that's what we should be talking about—whether determinism *really* has this bad consequence.

Frankfurt's alleged counterexample to the Principle of Alternate Possibilities is another example (Frankfurt 1969). In Frankfurt's story there is a mysterious character who, we are told, can prevent you from doing or deciding *anything* you might do or decide. But, in fact, he doesn't interfere with you because he happens to approve of what you do. In that case Frankfurt thought you must still be responsible because no one interfered with your doing what you wanted, even if you couldn't have done otherwise.

Frankfurt's story was supposed to undercut the traditional debate about whether determinism robs us of free will by convincing us that the ability to do otherwise isn't, after all, necessary for moral responsibility. His story didn't work, so his friends and supporters told other stories and an entire literature of "Frankfurt style-examples" sprung up and has lasted more than 40 years, with no signs of stopping or even slowing down. I have argued in print, more than once, (Vihvelin 2000 and Vihvelin 2008) that the stories don't work, that they are underdescribed thought experiments and that when you look more closely at the details, the subject of the stories never loses the ability to do otherwise. No one has said what's wrong with my argument. The response is always: "But, wait, here's another story."

However, it doesn't really matter whether I am right or wrong. The point is simpler. Frankfurt stories fail the two requirements on being a good thought experiment: that we all know what is being described and we agree about the verdict. A counterexample either works or it doesn't. If you have to spend 40 years arguing about whether a counterexample works, your thought experiment is a failure.

You might think, at this point, that this is just a problem for "Armchair Philosophy," and that the fix is to leave the armchair and to run some actual experiments.

But the questions that are used by the Experimental Philosophers include their own thought experiments, which are no less murky.

Here is a typical question (Nichols and Knobe 2007).

Imagine a universe (Universe A) in which everything that happens is **completely caused** by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, **if everything in this universe was exactly the same up until John made his decision,** then it ***had to happen*** that John would decide to have French Fries.

Now imagine a universe (Universe B) in which *almost* everything that happens is **completely caused** by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, **even if everything in the universe was exactly the same up until Mary made her decision**, it **did not have to happen** that Mary would decide to have French Fries. She ***could have decided to have something different***.

The key difference, then, is that in Universe A every decision is **completely caused** by what happened before the decision—**given the past, each decision *has to happen* the way that it does**. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision ***does not have to happen*** **the way that it does**.

Which of these universes do you think is most like ours? (circle one)

What is the question that is being asked? Do all the phrases used—"completely caused," "had to happen, given the past," and "could not have decided anything different"—mean the same thing? Do the people reading this questionnaire understand what these things mean? Are they all thinking about the same thing when they answer these questions? Are some of them perhaps thinking—as the move from "had to happen, given the past" to 'had to happen" and "could not have decided otherwise" suggests—of a world where there is no free will?

In the absence of any answers to these questions, these supposedly scientific thought experiments are no better than the Armchair variety.

Again, this isn't an argument against thought experiments in general. Nor is it necessarily an argument against thought experiments about free will. But the Rule, for free will philosophers, is this: Unless you know exactly what you are doing, and are sure you can do it well, avoid thought experiments (and avoid experimental philosophy).

Third Rule: Intuitions. Avoid them!

Or if you find that you cannot avoid them because all around you philosophers are appealing to intuitions, and arguing on the basis of their intuitions, and urging you to share their intuitions, constructing thought experiments designed to get you to have more intuitions, or writing up questionnaires for non-philosophers so we can have "data" about intuitions that are not confined to the intuitions of an elite group of philosophers, then I say "just don't do it."

Why not?

Because intuitions have no special evidential status, *qua intuition*. Why would anyone think they do? And how did all this talk of philosophical intuitions get going in the first place? This is relatively recent.

Intuitions are just a kind of belief and we don't think that beliefs *per se* have any special evidential status.

Again, I'm not saying it's *always* wrong to appeal to intuitions. Some kinds of belief are more epistemically trustworthy than others, and this may be true of some intuitions as well. Some people are very good at judging what other people are thinking and feeling simply by reading their faces and body language. Others—the autistic and aspergerish—are not so good. These kinds of intuitions don't have any philosophical payoff. But perhaps there are categories of philosophically relevant intuition which are highly reliable. One possible example might be beliefs about causation in particular cases. We have lots of daily experience of causation so *maybe* our intuitions about causation are a trustworthy source of data to constrain our philosophical theorizing.

But free will intuitions are very different from intuitions about causation.

In the case of causation, we have daily experience of particular cases that count as causation and cases that don't. We can tell the difference between one thing following another by co-incidence, and the first thing causing the second. In the case of free will, however, the clear contrast cases are few and far between. We have free will; rocks and plants don't. We are able to make choices we don't actually make and to do things we don't actually do. But beyond these clear starting points, things get confused and unclear very quickly. We all have free will—at least everyone in this room does. But when did we acquire it? At birth? When we learned to crawl, to talk, to ask questions, to argue with our parents? Or, as some of my students tell me, when we left home to go to university? Do we have free will all the time, or only some of the time? Do we have free will when we are asleep? Under the influence of alcohol or drugs? When we are in a state of panic or severely depressed? Do cats have free will? Might some form of artificial intelligence have free will? When I ask my students these questions, they tell me that they have never

thought about these things before, and many of them change their minds about the answers over the course of the semester.

When it comes to questions about free will and determinism, we have a positive reason to distrust our intuitions. Here's why. It's well known, in philosophy, that the fatalist is confused. Truth isn't the same as necessity, of any kind. The fact that there are truths about my future choices and actions does not affect my freedom *in any way*. But many years of trying to explain to my students why the fatalist is confused has convinced me that fatalist thinking runs deep. Some students get it; others never do. And it turns out that there are arguments for fatalism that are mistaken in ways that are much more subtle than the fatalist is usually given credit for.[2] So the situation is this: Even though it's a mistake, many people have the intuition that if it is "already true" what our future will be, then our future is not up to us; they think that truth alone—*regardless of determinism*—would rob us of free will. But if determinism is true, then there are detailed and specific truths about *all* our future choices and actions. So the intuition that determinism robs us of free will should not be trusted, for it might be a fatalist intuition in disguise.

Fourth Rule: Don't confuse "that's really strange" with that's impossible.

Compare for a moment, a very different literature—the literature about the possibility of time travel. Everyone in that literature understands that those who argue that time travel is impossible must show that the supposition that it is possible gives rise to *actual* contradictions (Lewis 1976). It is not enough to say—indeed, we can all agree—that a world where time travel takes place would be a most strange one.[3]

In the free will literature, by contrast, one often hears remarks to the effect that a deterministic world is a very strange one, and we would have to believe very strange—surprising!—things if we combine a belief that determinism is true with a belief that we have free will. For instance, we would have to believe that if I were to raise my arm just now, then either the remote past or the laws would be different.

But sometimes the surprising is true. This is what the history of science teaches us, and if philosophy is to make progress it should sometimes be what philosophy teaches us.

2. I argue this in Chapter 2 of Vihvelin 2013.

3. For argument that time travel is even stranger than Lewis thinks, see Vihvelin 1996.

Fifth Rule: Don't start with the hard cases.

Don't start with the cases where it isn't clear what to say because we don't know enough to know what to say or because we are confused or conflicted about what to say. The free will/determinism problem is the problem of deciding whether the truth of determinism would have the consequence that the Free Will fact is *never* a fact, not even in the easiest cases, the ones about which everyone agrees.

Sixth Rule: Don't analyze.

At least not at the start, not when you are defending the claim that we have free will (against someone who claims we never have it, or against someone who claims that having it is incompatible with determinism). If you proceed in this way, you are opening the door to the counterexample strategy. You are taking on a greater burden than you need to bear—the burden of defending the claim that your analysis gives the "correct" verdict in the hard cases as well as the easy ones.

Compare: You don't need an analysis of 'chair' or 'game' to be entitled to say that there really are chairs and games, nor do you need an analysis to have the right to say that the existence of chairs and games is compatible with determinism. Nor are you thereby committed to the claim that chairs and games are primitive components of reality.

Concluding Remarks

Back to the Free Will fact and the two objections that I mentioned at the beginning—the fatalist's objection and the incompatibilist's objection.

These objections are treated very differently in the current literature. It is almost universally assumed that the fatalist conclusion is wrong and that the only philosophical problem is to show what is wrong with the fatalist's arguments. (Not all of them are as obviously fallacious as the Fatalist Fallacy.)

But no one thinks this way about the hard determinist or the incompatibilist.

I blame this on the fact that argument by poorly described thought experiments and appeals to intuition is now widespread and common.

It wasn't always so. Back in 1983, back in the days when the incompatibilist was accused of making the kinds of mistakes the fatalist makes—of confusing causation with compulsion, descriptive with prescriptive laws, truth with necessity—Peter van Inwagen wrote an entire book (van Inwagen 1983) arguing that he, at least, was not guilty of any such simple mistakes. He claimed that there is an intuitively appealing and not obviously fallacious, argument for incompatibilism. He called it the Consequence argument.

I agree that this argument is not obviously fallacious. I also agree that the disagreement between us is not a merely verbal dispute. He asserts what I deny—that if determinism were true, then the Free Will fact would not be a fact. But, I claim, he is wrong. The Consequence argument fails. So far as *this* argument is concerned, it may be true that determinism is true and we have free will. And though I have devoted some time to this study, I know of no arguments that work.

So the state of play, at the present time, is that we have no reason to believe that the truth of determinism is incompatible with the Free Will fact. In the absence of other reasons—in the absence of some other *argument*—we are entitled to believe that the Free Will fact is a fact, and would be a fact even if determinism turned out to be true.

But I have digressed. I said that I would talk only about methodology, but I have ended up telling you the punch line of my book. I couldn't resist. But I still have free will. So I will exercise it by stopping.[4]

4. This paper is based on a talk presented at the Free Will Conference at the Center for Cognition and Neuroethics in Flint, Michigan on Oct. 11–12, 2014. Thanks to all who participated for their comments.

References

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 820–839.

Lewis, David. 1976. "The Paradoxes of Time Travel." *American Philosophical Quarterly* 13: 145–152.

Nichols, Shaun and Knobe, Joshua. 2007. "Moral Responsibility and the Cognitive Science of Folk Intuitions." *Noûs* 41 (4): 663–685.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Shoemaker, Sydney, 1969. "Time Without Change." *Journal of Philosophy* 66: 363–381.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Vihvelin, Kadri. 1996. "What Time Travelers Cannot Do." *Philosophical Studies* 81: 315–330.

Vihvelin, Kadri. 2000. "Freedom, Foreknowledge, and the Principle of Alternate Possibilities." *Canadian Journal of Philosophy* 30: 1–24.

Vihvelin, Kadri. 2008. "Foreknowledge, Frankfurt, and Ability to Do Otherwise: A Reply to Fischer." *Canadian Journal of Philosophy* 38 (3): 343–372.

Vihvelin, Kadri. 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.

cognethic.org