



Journal of Cognition and Neuroethics

ISSN: 2166-5087

April, 2014. Volume 2, Issue 1.

Journal of Cognition and Neuroethics

Managing Editor

Jami L. Anderson

Production Editor

Zea Miller

Publication Details

Volume 2, Issue 1 was digitally published in April of 2014 from Flint, Michigan, under ISSN 2166-5087.

© 2014 Center for Cognition and Neuroethics

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

Introduction

Jami L. Anderson

- | | | |
|---|--|---------|
| 1 | Is Reason Contradictory When Applied to Metaphysical Questions?
Graham Schuster | 1–11 |
| 2 | Implicitly Grounded Beliefs
Andrew Koehl | 13–36 |
| 3 | The Enigma Of Probability
Nick Ergodos | 37–71 |
| 4 | The View from Vector Space: An Account of Conceptual Geography
Joshua Stein | 73–93 |
| 5 | The Self-Awareness of Reason in Plato
Daniel Bloom | 95–103 |
| 6 | Reasoning with and without Reasons: The Effects of Professional Culture and Information Access in Educational and Clinical Settings
Barry Saferstein | 105–125 |
| 7 | Moral Heuristics and Biases
Mark Hermann | 127–142 |
| 8 | Reasoning and the Military Decision Making Process
Ibanga B. Ikpe | 143–160 |

9	The Role of Emotional Intuitions in Moral Judgments and Decisions Catherine Gee	161–171
10	Asking for Reasons as a Weapon: Epistemic Justification and the Loss of Knowledge Ian Werkheiser	173–190
11	Brain Rays, Advertising, and Fancy Suits: The Ethics of Mind Control Brent Kious	191–210
12	Philosophy and Neurobiology: towards a Hegelian Contribution on the Question of the Juridical Status of the Human Embryo Fernando Huesca Ramón	211–220

Introduction

Jami L. Anderson

University of Michigan-Flint

We are pleased to announce the second issue of the *Journal of Cognition and Neuroethics* (JCN). The primary goal of JCN is to provide an open-access forum in which scholars from a wide variety of backgrounds, fields, professions, and disciplines can contribute to discussions concerning cognition and neuroethics. The papers included in this second issue are revised and expanded versions of papers presented at the first annual Center for Cognition and Neuroethics conference, Reason, Reasons and Reasoning which was held October 11–12, 2013. The presenters, who came from across the country and from around the world, were from a variety of professions, institutions and subject disciplines. Some of the papers presented at that conference have been included in this issue. While a few papers included here address very abstract puzzles concerning reason, others examine questions about reason in concrete, practical terms. Some examine implications of fallacious reasoning patterns while others consider ethical problems raised by various reasoning processes.

CCN was formed with the goal of fostering genuinely interdisciplinary research with the belief that the skills and knowledge that individuals in different fields and whose work relies on or assume different assumptions can enrich and inform one another's work. This issue, the outgrowth of the efforts of individuals from a variety of professional backgrounds, is a demonstration of the value of interdisciplinary projects.

As always, the time and efforts of our manuscript reviewers is very much appreciated. For more on the *Journal of Cognition and Neuroethics* or the Center for Cognition and Neuroethics, we invite you to visit our website at cognethic.org.

Jami L. Anderson

Managing Editor

Journal of Cognition and Neuroethics

Journal of Cognition and Neuroethics

Is Reason Contradictory When Applied to Metaphysical Questions?

Graham Schuster
University of Georgia

Biography

Graham Schuster is a Ph.D. candidate at the University of Georgia, where he is working on Hegel's attempt to address the possibility of unconditional knowledge. He is interested in any serious inquiry into the nature of what is, with a particular focus on Plato, Aristotle, the early Moderns, Kant, and Hegel. He is also considering the fundamental assumptions of Sartre and Nietzsche as part of an investigation of the role of creativity and judgments of beauty in moral engagement with the world. Mr. Schuster is an adjunct professor at Oglethorpe University, where he teaches a series of interdisciplinary courses in existentialism as an investigation into human freedom, using both philosophical and literary source material.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Schuster, Graham. 2014. "Is Reason Contradictory When Applied to Metaphysical Questions?" *Journal of Cognition and Neuroethics* 2 (1): 1–11.

Is Reason Contradictory When Applied to Metaphysical Questions?

Graham Schuster

Abstract

With his “Antinomies of Pure Reason,” Kant exposes the illusory nature of reason’s attempts to think the unconditioned in cosmological, metaphysical questions. The illusion is that reason can have anything to say at all on these matters. When one tries to put forward an argument that, for example, the world must have a beginning in time, one finds that there is also an argument ready to hand that the world must be infinite in temporal extent. Since we cannot rest in a contradiction, the result is that reason comes to recognize the boundaries of its own appropriate use. This crucial step in Kant’s project requires that reason produce equally convincing arguments with opposite conclusions. This paper argues that Kant’s Antinomies are, however, not convincing, because they engage in question-begging. This is not, however, simply a mistake on Kant’s part. Rather, it follows from his fundamental assumption that concepts are fixed in character. If, however, concepts of pure reason have within themselves sufficient determination for a dynamic self-development, then the way is open for a rational grasp of the unconditioned that does not involve question-begging, nor unavoidable contradictions.

Keywords

Kant, Critique of Pure Reason, Hegel, metaphysics, epistemology, reason, reasoning, antinomy, concept, idealism, contradiction, time, cosmology, unconditioned, conditioned, appearance

All pre-Kantian attempts to do metaphysics, that is, to attain knowledge of the nature of reality, fail, and all for the same reason—they assume that knowing consists of one’s ideas matching the independently given objects of those ideas. The trouble is, one can never get outside of one’s ideas to verify their supposed correspondence with their objects. This leads the pre-critical philosopher either to the enthusiastic but unverifiable conviction that one’s ideas can and do match objects, or to the skeptical recognition that one’s experiences of objects carry no necessity, and at best one’s ideas can be shown to correspond to each other, but not necessarily to a mind-independent world.¹ If this fundamental assumption about knowledge were unavoidable, we could conclude that metaphysics is impossible. Kant, however, offers another option. If we invert the relationship between knowing and its objects, such that knowing is constitutive of its

1. Locke is a good example of the former, Hume of the latter. See also Kant, *Critique of Pure Reason*, B127-28 (hereafter cited KrV).

objects, then by knowing our own knowing, to which we do have access, we will at the same time be able to know the objects that our knowing produces. In what follows, I hope to show that Kant's account would preclude the possibility of a genuine metaphysics, but that one of his key arguments, that of the Antinomies, fails in its aim. Further, I hope to show that it must fail, but that seeing why it must fail points the way to a genuine metaphysics, such as what we find in Hegel's work.

I

Starting with the, initially hypothetical, assumption that knowing is constitutive of its objects, our task is to find the elements of knowing that are necessary for any awareness of an object. These elements are (1) the *a priori* intuitions of space and time and (2) the concepts of the understanding (categories). One could summarize the bulk of the first half of the *Critique* by saying that without these three factors at work in us *a priori*, no experience of an object would be possible. As such, they are features of our awareness that are constitutive of objectivity in general.

Spatial determination is necessary for any awareness of an external object because, minimally speaking, such an object must be represented as outside of oneself. If space belongs to things, then one would first have to experience things, then abstract one's representation of space from these experiences. But, spatial representation is a prerequisite for the first experience, so space belongs to our way of perceiving, not to a mind-independent world (*KrV* A23/B38, A26/B42).

Likewise, temporal determination is necessary for the awareness of any object because, minimally speaking, an object must be represented as existing simultaneously with oneself, and coming to be aware or ceasing to be aware of an object requires the representation of succession. Also, one's awareness itself occurs as a series of changing inner states that belong to one, persistent self. Temporal determination is thus a necessary condition of inner experience, and since the awareness of objects occurs in inner experience, time must belong to us, not to things. If time belongs to things, then just as with space, one would have to experience things first, and then abstract the representation of time from those experiences. But one cannot have any experiences without temporal determination already at work making the experience possible (*KrV* A30/B46).

In addition to space and time as the conditions of sensibility, conceptual determination is necessary for the awareness of any object because, as Kant famously says, intuitions without concepts are blind (*KrV* A51/B75). That is, the manifold of sensible content that we are aware of as temporally and spatially determined is, without concepts, just a

meaningless jumble of data. Concepts allow us to unify the manifold stream of incoming data in various ways that make it intelligible. For example, without the concept of thinghood (substance), one could not be aware that some particular collection of sensible content belongs to one thing as distinct from other things. Without the concept of cause and effect, one could not be aware of the connection between one state of affairs and the next, so one could not experience an event. And so on.

Furthermore, since the categories make objectivity in general possible, they must apply to all objects, which means they cannot have any particular content of their own. They are dependent on a matter given in sensibility. As such, their legitimate use extends only to that which can be given in sensibility, that is, to appearances.

Of course, there is much more to the story, but for my present purposes, the upshot of all this is that space, time, and concepts—the elements of human knowing—must be found in every experience, since they are constitutive ingredients of experience in general. But at the same time, the *way* things appear (spatially and/or temporally) and *what* things appear as (their conceptual determination) pertain only to how things are *for us*, as possible objects of experience, but are utterly meaningless if ascribed to things in themselves, independent of our awareness of them. This is analogous to, for instance, the sweetness of a fig. “The fig itself is sweet” is really a meaningless statement, since the sweetness belongs to our perceptual faculties. As conditions for the possibility of objects, the fundamental elements of our experience of objectivity cannot belong to the objects, since the condition of something is logically prior to that which it conditions. Space, time, and concepts are, in Kant’s terminology, transcendently ideal. They are merely ideal, as opposed to real, when applied outside the boundaries of possible experience.

If Kant is right that what we can know of objects is what we put into them, then the nature of our perceptual and conceptual determinations means that our knowledge is limited to appearances. Although it counts as real knowledge, because it pertains to how things *must* appear, it does not extend to the way things are in themselves. In other words, since our knowing conditions what can be an object for us, any possible object of knowledge is necessarily conditioned, and thus, we cannot know that which is unconditioned. As we will see, this gets reason into trouble.

II

Reason finds itself confronting what Kant calls an illusion, the source of which, generally speaking, is the ascription of objectivity to that which cannot be an object (KrV A297/B353). In other words, reason cannot help but take its own principles and

apply them as if to objects, yet these so-called objects are entirely outside the scope of a possible experience, and hence, not really objects at all.

The illusion, Kant argues, is built into the very structure of reason, which compels it to seek the unconditioned, regardless of the particular contents of its thinking. For, reason draws inferences, that is, it justifies some judgment (the conclusion) by means of its connection with other judgments (the premises). Reason is thus properly called a faculty of syllogizing, the activity of which is to uncover, and provide to thinking, the conditions for something that is thereby conditioned. The nature of this activity, however, compels reason always to seek yet higher conditions, because the premises in any given instance of reasoning are themselves judgments that require justification. This feature of reason is not by itself illusory, although it does involve the difficulty associated with a regress of conditions, such that no syllogism can ever be completely grounded. It may well be the case that there is a regress here that reason cannot complete. This would be unsatisfying in terms of the possibility of reason leading to unconditional knowledge, but it is not illusory. The illusion becomes evident by adding, to the unavoidable regress of conditions, another necessary feature of reason, namely, the Principle of Sufficient Reason. That is, for any given conditioned term, since its determination depends on its conditions, all of its conditions must also be given (*KrV* A307/B364). If this is correct, and if some conditioned term is given, then the principle amounts to positing an unconditioned, either as the totality of the regressive series, or as standing somehow at the head of the series. The totality would be unconditioned because, as the complete series of conditions, there can be no condition outside of it. And a first member would be unconditioned, because if it had a condition, it would not be first.

In whichever of these two ways the unconditioned is conceived, reason takes it as something objective, that is, as having membership in the series of conditions. For, although reason has its own principles, it has nothing upon which to operate other than judgments of the understanding, which it connects into a unity that is yet itself a judgment. Accordingly, although reason does not refer directly to objects, its matter—judgments of the understanding—applies only to objects. Hence, reason cannot help but take its inferences to have objective significance. Since reason seeks the unconditioned as the ultimate ground of the conditioned, and indeed provides the principle whereby this relation is necessary, reason is engaged in uncovering the ultimate condition(s) of appearances. In so doing, reason not only claims objective significance for the unconditioned, but as we have already seen, objectivity is the conditioned result of the activity of sensibility and understanding (intuitions and concepts).

To reiterate, the empirical demand that everything has a condition does not of itself lead to the unconditioned. Rather, it simply leads to a regress of conditions. (*KrV* A308/B364) However, reason's own demand for a complete totality of conditions leads it to go beyond what is merely given empirically, and thus posit, as an actual object, the unconditioned as standing in a necessary relationship to the series of appearances. Thus, it is a condition of reasoning that the unconditioned be conditioned, even though this is an impossibility. Hence, although one sees the conflict, one cannot escape the illusion without abandoning the use of reason. However, one cannot abandon reason, because one cannot abandon seeking the unconditioned basis of the conditioned, as is evident in Kant's examples of its use. Thus, we cannot help but reason about the unconditioned, but when we do, we contradict ourselves, as Kant argues in the antinomies.

III

The antinomies are supposed to be both illustrations of, and antidotes to, the illusion that reason finds itself in when it oversteps its boundaries and tries to ask metaphysical, cosmological questions. They aim to show that reason finds itself not only in an illusion, but in direct conflict with itself when it attempts to think the unconditioned. This is because there are two ways to think the unconditioned in the series of appearances. That is, the unconditioned could be some condition that is itself unconditioned, that is, something standing at the head, or outside, of the series of conditions, but not itself under any further condition. Or, the unconditioned could be the totality of conditions, that is, the series itself, outside of which there can be no condition, since any condition must be included in the sum total of all conditions (*KrV* A417/B445). The trouble is, in the case of the first two antinomies, the unconditioned has to be one or the other, but both options are impossible. The only way out of the conflict is to deny reason the ability to ask such questions. If successful, each antinomy amounts to an argument for the restriction of reason to the domain of appearances.

The first antinomy deals with the extent of the world in space and time.² The argument of the thesis is as follows. Assume that the world never began, that it is infinite in temporal extent. Then, because the series of prior times must be completed to arrive at the present, an infinite series of prior times must have been completed. But, by definition, an infinite series cannot be completed. Therefore, there is no present. Since we

2. For the sake of brevity, I am only going to address the time component of the first antinomy, as an example of what I think applies to the antinomies in general.

have clearly arrived at the present, the initial assumption must be false, and the world is finite in temporal extent (*KrV* A426/B454).

The argument of the antithesis is as follows. Assume that the world began, that it is finite in temporal extent. Then, there must have been an empty time before the world began. In an empty time, there is nothing that could act as a cause, since causes and the laws they follow belong to the world. But, for the world to begin, something has to cause it to begin. Therefore, the world did not begin, and does not exist. Since the world clearly exists, the initial assumption must be false, and the world is infinite in temporal extent (*KrV* A427/B455).

Both arguments are valid and, Kant thinks, convincing. But their conclusions cannot both be true. The world cannot be both finite and infinite. So Kant offers a simple solution; the world is *neither* finite nor infinite, or rather, we can make no meaningful claim about it one way or the other. In order to make these contradictory arguments, one has to proceed on the underlying assumption that the resources available to reason are adequate to a grasp of the unconditioned. The thesis turns on the claim that an infinite world would involve a successive temporal series that cannot be completed. But notice that this places the infinite world-whole under the conditions of temporal succession that govern the conditioned entities within the world. In other words, in trying to grasp the world as an unconditioned, infinite totality, one can only think of it in terms of the temporal conditions of sensible appearance, and thus one fails to grasp the world as an unconditioned totality.

A similar assumption haunts the antithesis, but in the other direction. There, one takes a member of the series, which as such stands under the temporal conditions of sensible appearance, and attempts to think it as not under these conditions, namely, as not in a necessary causal connection with a prior state of affairs. But, membership in the world-series involves the determinations of natural law, which are necessarily temporal, for instance, the requirement that every event have a prior cause. This means that, in trying to think a member of the series as unconditioned, one fails to think a member of the series.

What generates the conflict, Kant thinks, is the uncritical use of reason. What the antinomies have in common is the underlying assumption that perceptual and conceptual determination apply outside of what can be an object of experience. These arguments are put forward by a reason that has failed to discover that intuitions and concepts apply only to appearances. As soon as we notice that the unconditioned cannot be an appearance, and that temporal determinations belong only to appearances, then it becomes clear

that it is simply untrue to say that the world as unconditioned whole, or an absolute beginning of the world, is conditioned by any temporal determination whatsoever.

Thus, there is really no conflict here. Both thesis and antithesis are false. Since the only way out of the dilemma is to pull reason back from its illegitimate transcendental realism, the antinomies turn out to function as an argument for transcendental idealism, for the limitation of reason to the realm of appearance, which is to deny it any satisfaction in the pursuit of metaphysical questions.

IV

I think, along with Hegel, that we need not accept Kant's conclusion limiting the scope of reason. The antinomies are not convincing because their arguments assume what they are trying to prove. The question-begging takes place at two levels, with respect to (1) the individual arguments and (2) the overall argument for limiting reason.³ In the thesis, the question-begging can be seen in the premise that an infinite series cannot be completed. By hypothesis, now exists and the series leading to it is infinite. If we assert that an infinite series cannot be completed, then we have already denied the hypothesis, but the denial of an infinite series leading to now is just what the argument seeks to prove. The argument is valid, but trivial. It amounts to saying that if there can be no infinite series, then there can be no infinite series.

Regarding the antithesis, we have the claim that any state of the world must have a cause. Yet, by hypothesis, there is an uncaused beginning state of the world. As soon as we assert the claim, we deny the hypothesis, which denial has yet to be proved. Again, the argument is the equivalent of saying that if there can be no unconditioned beginning, then there can be no unconditioned beginning.

If I am correct that the antinomies beg the question,⁴ then noticing the reason why they do so could be instructive in terms of the possible application of reason to metaphysical questions in general. Given the result of Kant's account, which precedes the antinomies, that our cognitive apparatus only has relevance to that which is conditioned as an appearance, the antinomies cannot help but beg the question. This is because, lacking any insight into the unconditioned, reason has no resources available to generate meaningful premises regarding the essential attributes of the unconditioned. To avoid

3. For my present purposes, I ignore Hegel's treatment, in favor of my own simpler argument, but see G.W.F. Hegel, *Encyclopaedia Logic* §48; G.W.F. Hegel, *Science of Logic*, 190–97. See also Sedgewick (1991).

4. Although I only treat the first antinomy here, the others fall prey to a similar critique. Indeed, any argument concerning the unconditioned must be circular, given transcendental idealism.

begging the question in the thesis, for example, one would need to justify the claim that an infinite world-series cannot be completed. But, since an infinite world-series, as unconditioned, cannot be an object of knowledge, one cannot defend the claim, and is left with no option other than mere assertion. In other words, lacking insight into things in themselves, any claim about a thing in itself can only be asserted without a rational basis.

Not only do the individual arguments beg the question, but the antinomies in general do. Kant's goal is to show that transcendental idealism is correct, but the whole endeavor of the antinomies assumes that any thinking about an object is transcendently ideal. For the hypotheses to be shown as absurd, one has to assume that temporal determination in general is equivalent to the temporal determination of appearances, that the concept of time cannot be reasoned about outside of the conditions of possible experience. So, in the thesis, one can maintain that an infinite world-series cannot be completed only on the assumption that "time" is equivalent to "time as experienced," namely as a successive synthesis of moments. Of course, one cannot traverse an infinite series of times in experience, but why should we accept that time itself must be traversed in order to be given? Likewise, in the antithesis, one can maintain that every time must be determined by a prior time only on the assumption that "an individual time" means "an individual time as experienced," which necessarily involves a causal connection with a prior time.

Only a transcendental idealist would accept these assumptions, yet the antinomies are supposed to argue for transcendental idealism. Thus, the both the content and the overall strategy of the antinomies commits Kant to begging the question.

V

Where does this leave us with regard to the possibility of reasoning about metaphysical questions? The assumption of transcendental idealism forces the individual arguments into circularity and itself amounts to question-begging, since the antinomies are meant to prove transcendental idealism. Kant makes this assumption going in to the antinomies because it is the result of the first part of the *Critique*. So, the antinomies' failure indicates that there might be something wrong with the preceding account. As it turns out, there is a fundamental assumption that underlies Kant's entire account, and which leads directly to the transcendental idealist position. I suggest, along with Hegel, that the problematic assumption is that concepts are fixed in character.

Concepts, according to Kant, are entirely dependent for content on a matter given by sensible intuition. Since the *a priori* concepts make objects in general possible, they

must apply to any object, which means that they have to be indifferent to the particular contents that distinguish one object from another. In other words, concepts, considered apart from sensible intuition, are empty and have no resources of their own to generate any content. They are empty, fixed universals (*KrV* A51/B75). But this, in turn, means that conceptual determination has no significance without sensible content, that is, concepts only apply within the domain of appearances.

With this assumption about the nature of concepts at work, that they must apply to any object and are thus empty, it is no surprise that reason cannot generate nontrivial premises regarding the temporal extent of the world. For, this would require one to think the concepts “finite” and “infinite” as having significance for an unconditioned world-whole or absolute beginning. On Kant’s assumption, however, these concepts can only have significant application to a sensibly given content. And so, the antinomies illegitimately apply the conditions of temporal appearances to that which, by hypothesis, is unconditioned.

If the problems here are generated by the assumption of the fixed nature of concepts, then it might well be the case that reason is not in conflict with itself in asking metaphysical questions, but, rather, in conflict with this assumption. We are now in a position to question the conclusion of the antinomies, but in a way that indicates the next move. We can, at least provisionally, go forward with the supposition that concepts can have intrinsic content. If this is correct, then reasoning using concepts alone, that is, philosophy, could think the unconditioned without inadvertently rendering it conditioned, and thereby contradicting itself. This is essentially Hegel’s move. For the post-Kant development of a metaphysics that does not get stuck in contradiction, nor in circularity, I refer you to Hegel’s *Science of Logic*, which opens the possibility of thinking the unconditioned by showing how conceptual determinations develop out of their own inner necessity.

References

- Hegel, G.W.F. 1969. *Hegel's Science of Logic*. Translated by A.V. Miller. London: George Allen & Unwin.
- Hegel, G.W.F. 1975. *Logic: Being Part One of the Encyclopaedia of the Philosophical Sciences*. Translated by W. Wallace. Oxford: Clarendon Press.
- Kant, Immanuel. 1998. *Critique of Pure Reason*. Translated by P. Guyer and A.W. Wood. New York: Cambridge University Press.
- Sedgewick, Sally. 1991. "Hegel's Strategy and Critique of Kant's Mathematical Antinomies." *History of Philosophy Quarterly* 8 (4): 423–440.
- Sedgewick, Sally. 1991. "Hegel on Kant's Antinomies and Distinction Between General and Transcendental Logic." *The Monist* 74 (3): 403–420.

Journal of Cognition and Neuroethics

Implicitly Grounded Beliefs

Andrew Koehl

Roberts Wesleyan College

Biography

Andrew Koehl is Professor of Philosophy at Roberts Wesleyan College. He received his Ph.D. in philosophy from the University of Notre Dame, with specializations in historical philosophy, epistemology, philosophy of religion, and ethics. His research interests include epistemology, religious diversity, and applications of Aristotelian ethics.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Koehl, Andrew. 2014. "Implicitly Grounded Beliefs." *Journal of Cognition and Neuroethics* 2 (1): 13–36.

Implicitly Grounded Beliefs

Andrew Koehl

Abstract

In 1945, W.T. Stace addressed the epistemic status of the “unreasoned belief,” one that “is neither a case of immediate knowledge nor has been reached by a process of reasoning.” In this paper I call such beliefs “implicitly grounded beliefs” (IGBs) because, as Stace acknowledged, their commonality lies not in their being unreasoned or unreasonable, but in the implicitness of their grounds. Stace argued that if philosophers employ implicitly grounded beliefs in their work, they must try to reconstruct the inexplicit processes that led to them. This may be so, but the question of how implicitly grounded beliefs should be treated in academic works is separate from the question of whether or not such beliefs can be justified, warranted, and known. Here I first explore the nature of implicitly grounded beliefs and then—drawing upon the most common internalist, externalist, and aretaic conceptions of positive epistemic status—I argue that implicitly grounded beliefs can indeed be justified, warranted, and known.

Keywords

Epistemology, implicitly grounded beliefs, intuition, experts, W.T. Stace, warrant, justification, internalism, externalism, aretaism, special access, perspectively basic beliefs

In 1945, W.T. Stace addressed the epistemic status of the “unreasoned belief,” one that “is neither a case of immediate knowledge nor has been reached by a process of reasoning” (1945a, 29). In this paper, I call such beliefs “implicitly grounded beliefs” (IGBs) because, as Stace acknowledged, their commonality lies not in their being unreasoned or unreasonable, but in the implicitness of their grounds:

Except perhaps in the case of beliefs produced by pure random conditioning, the so-called unreasoned beliefs only *appear* to be unreasoned, and the psychological processes only *appear* to be non-logical. Most of them are in their essence reasoning processes, though the reasoning may be more or less crude. We are not aware of their rational character, or even of their existence, because of their unexplicitness. (1945b, 141)

Stace bemoaned the fact that philosophers appeal to unreasoned beliefs in their work:

Can we say anything which could put a stop to the present practice of treating [unreasoned beliefs] all alike as final authorities, without any discrimination of their respective merits, so that philosophy is simply a battleground of final authorities which contradict one another? (1945b, 143)

Stace argued that if philosophers employ implicitly grounded beliefs in their work, they must try to reconstruct the inexplicit processes that led to them. But the question of how implicitly grounded beliefs should be treated in academic works is separate from the question of whether or not such beliefs can be justified, warranted, and known. After exploring the nature of implicitly grounded beliefs below, I will argue that they can be.

I. Implicitly Grounded Beliefs – An Introduction

A. Formed by inexplicit reasoning processes

Scott Johnston, founder of Sterling Financial Group, consistently outperformed other well-trained and experienced money managers, even though his strategies were similar to those of managers who fail. When asked whether there was some other factor behind his success, Johnston responded,

Yes, there is. It's that the very best managers develop a sixth sense where they just know that a stock is going to move...You develop a sixth sense, an instinct. We're talking art here, not science. Many have the ability, the training, the commitment, but few have the touch... It's visceral. You just sense it. You know that a stock's got all the elements to be a winner. It just feels right; it's ready to move. (Tannous 1997, 163)

Every field harbors remarkable people who seem to have amazing, intuitive ways of knowing. These experts seem to arrive at justified and warranted beliefs, even though they cannot specify all of their grounds or methodologies. There are chess masters who know just the right move to make while playing twenty opponents simultaneously, interviewers with a knack for evaluating job candidates, fishermen who have a “feel” for “reading a river,” detectives with an uncanny insight for solving crimes, nurses who implicitly pick up on subtle cues to conclude that a newborn is in trouble, social workers who know exactly what kind of help will work for a client, performance artists who

intuitively know when the audience is ready for a bit of humor, interpretive dancers who can just sense the best time to pause, mathematical prodigies who perform incredible calculations instantly, and parents who have an intuitive grasp of a child's state of mind that outsiders do not possess.

In one survey, seventy-two of eighty-three Nobel laureates in science and medicine implicated intuition in their success. 'We felt at times that there was almost a hand guiding us,' said Michael Brown, winner of the 1985 prize for medicine. 'We would go from one step to the next, and somehow we would know which was the right way to go, and I really can't tell how we knew that.' (Myers 2002, 61)

Experts are often much better at forming true beliefs in their areas of expertise than they are at verbalizing the grounds for their beliefs or at identifying the methods they employ to arrive at their beliefs (Speelman 1998. See also Griffin, Schwartz and Sofronof 1998, 333–337.). Often, when experts try to teach their skill to novices, they create a "reasonable" guess at their methodology that does not match how they actually achieve their remarkable results (Speelman 1998, 136). Rouse and Morris show that experts may employ methodologies that are not easily accessible to verbalization, such as "conceptually abstract, pattern-oriented mental models" (Rouse and Morris 1986, 145). Each of us has some area of intuitive expertise, and most of us, to some degree, can discern when another person is angry, frightened, deceptive, or trustworthy even when we cannot explicitly identify the aural or visual cues that create such convictions in us. On a more basic level, our understanding of language and grammar is mostly implicit:

An example: You know which of these two phrases sounds better: 'a big red barn' or 'a red big barn' – but your conscious mind struggles to articulate the rule you intuitively know. (Myers 2002, 52)

Even if we do not think of ourselves as particularly intuitive, we rely on implicit modes of belief-formation throughout each day of our lives.

IGBs do not arise from any special non-natural powers; they are not cases of extra-sensory perception or anything paranormal. The inputs to IGB-formation processes stem from natural sources, such as sense perception, stored memories, introspection, or reasoning (Szalita-Pemow 1955). As Stace points out, the main inexplicit reasoning processes seem to be varieties of the association of ideas (according to similarity and analogy, coherence, stability, esthetic or conceptual fittingness, etc.) (1945b, 140). In recent years the dual-process model of cognition has become ascendant (See Evans and

Frankish 2009, Barrett, Tugade and Engle 2004, and Strack and Deutsch 2004.). The System 1 processes are the “automatic, associative, nonconscious processes” (Kaufman 2009, 28) that lead to IGBs, as opposed to System 2, which involves “controlled, deliberate, reflective processes” (Kaufman 2009, 28). Others see this model as too general. For instance, Gore and Sadler-Smith delineate four primary types of intuition (what I am calling IGBs): problem-solving, creative, social, and moral. Each type of IGB comes about through distinct processes, and involves the activation of different parts of the brain (Gore and Sadler-Smith 2011). Some IGBs may arise from the process of chunking, where a thinker accrues a large number of perceptual patterns, called *chunks*, which are “a collection of elements having strong associations with one another, but weak associations with elements within other chunks” (Gobet, et al. 2001, 236). These chunks may be organized into larger patterns, or *templates*.¹ Some intuitive people may employ largely spatial or pictorial mental models, which helps to explain why they may be unable to verbalize the grounds for their beliefs (Rouse and Morris 1986). Each mental process may be impacted by desires, moods, emotions, personality (optimism or pessimism; attitude towards risk), and cognitive style. Implicitly grounded beliefs may flash into consciousness like an *a priori* intuition, but they are actually the result of previous experience and thought. The effect of previous learning on present thought, even apart from conscious remembrance of that learning, has been well documented.² Past experiences and reasoning provide the materials for inexplicit belief-formation processes.³

It is important to emphasize that IGBs are not guesses. Implicitly grounded beliefs differ from guesses in three ways. First, a guess comes about when one has nothing to go on. For example, I meet someone I’ve never seen before who asks me which of his

1. “Templates, which are a special kind of chunk, possess both a core, made of stable information, and slots, made of variable information.” Chassy and Gobet 2011, 200.

2. For instance, R. Crutchfield (Crutchfield 1960) demonstrated that a subject’s ability to solve puzzles increased significantly if she had previous experience solving similar puzzles. He gave subjects spatial organization puzzles to solve which involved principles some of which were pertinent to later puzzles. He found that subjects who had previous experience with puzzles exhibiting relevant principles were much more successful in solving those puzzles, even though the subjects themselves could not recall any of the principles nor did they recall conscious use of them. For more discussion of the implicit influence of previous experience on present thought, see Eyesenck 1990, Schachter 1987, Myers 2002, Kunst-Wilson and Zajonc 1980, and Graf and Masson 1993.

3. Among the many philosophers, psychologists, mathematicians, and scientists who have explored or employed the notion of inexplicit reasoning processes as the source of implicitly grounded beliefs are Berne 1949, Berne 1962, Bunge 1962, Hadamard 1954, Mach 1896, Poincare 1946, Russell 1948, Russell 1956, Thrash and Elliot 2003, and Wallas 1926.

pockets contains a coin. Since I do not have any previous experience that would give me an insight into his coin-placing tendencies, I randomly choose. I guess. Conversely, implicitly grounded beliefs arise from relevant past experience and/or background beliefs and impressions that provide insight. Such relevant resources are unavailable in the case of a mere guess. Consequently, IGBs are true in significantly more cases than guesses. Finally, an IGB is often a firm conviction that a proposition is true, not an unmotivated choice. It is a kind of *belief*, whereas a guess is not.

While cognitive psychologists usually call beliefs formed in this implicit sort of way “intuitions,” in epistemological circles the term “intuition” usually indicates an *a priori* belief that concerns only abstract and necessary propositions—such as the intuition that if all A are B, and all B are C, then all A are C. IGBs are *a posteriori*, which means that one must have specific previous kinds of experience and evidence in order to form them, and they can concern all manner of propositions, including concrete and contingent propositions. When implicitly grounded beliefs do concern matters of necessity, they lack the immediacy and force of a rational intuition. Rational intuitions according to epistemologists are such that merely understanding the proposition compels belief, and yet merely understanding a proposition cannot produce an IGB. Absent evidence from past experiences, no IGB will be formed in response to a proposition like “Apple Computer stock is about to move.” For instance, when I read Professor Phelps’ proposed sufficient conditions for knowledge I think, “That can’t be right.” Only upon ample reflection do I conceive of a clear counter-example. At first I have an IGB that his conditions for knowledge are inadequate, and only later do I have the rational intuition. At that point I consider the counterexample and can “just see” that Phelps’ conditions are not sufficient.

Since I am arguing for an epistemological thesis in this paper—that such beliefs can be justified and known—I will label them not *intuitions* but *implicitly grounded beliefs (IGBs)*. (I will retain the term “intuitive” and will use it to reference those who form IGBs.) Implicitly grounded beliefs arise from thought processes not explicitly recognizable to the believer. They are sometimes basic (that is, not inferred from any other beliefs), but they are always *perspectivally* basic: The subject cannot make explicit the actual grounds of her belief, nor can she manufacture a good justifying argument for her belief. Sometimes the subject can make *some* of her grounds explicit, but not enough to provide justification for her belief.

B. Distinguished from other categories of belief

Other categories of belief are similar to rational intuitions, and each of them, like rational intuitions, is distinguishable from IGBs. These include *categorical and metaphysical intuitions*—such as the intuition of space-time, of causality, of the independent existence of objects, or of the existence of other persons; *moral intuitions*—such as the intuition that we should sometimes sacrifice our own well-being for the good of others; *mystical intuitions*—such as Bergson’s intuition into a trans-categorical unity of all events and objects, Plato’s apprehension of The Good, Heraclitus’ intuition of the *logos*, and mystical religious apprehensions of God; *paranormal deliverances*—such as beliefs formed through ESP, clairvoyance, or past-life memories.

Each of these categories of belief differ from implicitly grounded beliefs in that, like rational intuitions, they are taken to be immediate and direct apprehensions, rather than the conclusion of inexplicit reasoning processes drawing upon information gained through past experience. The grounds for the above mentioned beliefs are taken to be entirely explicit and occurrent. Another difference is that these beliefs, unlike IGBs, are usually thought to have a special status such as being *self-evident* or *infallible*. Likewise, in most of the cases cited above the insight in question could not even in principle be arrived at through one’s other faculties, but this is not the case with IGBs.

IGBs clearly differ from perceptual beliefs, memory beliefs, and introspective beliefs as well. The most obvious difference between *perceptions* and IGBs is that perceptions only concern sensory matters whereas IGBs can concern a wide array of propositions. Further, perceptual beliefs are differentially sensitive to what transpires in the field of experience, whereas nothing analogous is true of IGBs.⁴ An *introspection* makes evident a proposition about my own mental states, whereas IGBs concern a wide range of facts external to

4. There are some similarities, however. When I perceive Lydia what I perceive are not just shapes and colors or even just a person, but I see what I see as Lydia. My perception of Lydia comes about not only through current sensory input, but through a linking of current input to stored experiential and conceptual information. My perception of Lydia is a combination of a seeing and an inexplicit thought process. A similar process produces the IGB that Lydia is lying. As I interact with Lydia I observe (perhaps without explicitly taking note of them) certain cues like a shaking hand, dilated eyes, a slight waver in her voice, or perhaps a slightly strange tone or cadence in her way of speaking, all elements which in my past experience, perhaps completely unreflectively, I’ve noticed to be signs of dishonesty. As I look at Lydia now, without even consciously recognizing the particular signs, it seems unmistakably to me that she is lying.

The difference between such an IGB and a perception is that while this IGB involves current sensory input, it is not differentially sensitive to the field of experience as sensory experiences are. Even if I didn’t recognize the person I saw coming towards me as Lydia, and thus didn’t have the perception of Lydia, the sensory experience would still be intricately informative. I would recognize the person I saw as a woman of medium height,

me. Also, introspections, unlike IGBs, are thought to be immediate and by some to be infallible. Finally, IGBs are different from *episodic memories*, in that episodic memories are always autobiographical - they cause one to “re-live” parts of one’s own past—and they usually involve more vivid imagery than IGBs.⁵ However, *semantic* memory—the remembrance of facts—is a kind of implicitly grounded belief, at least when one cannot remember the sources of one’s belief.

C. IGBs Mistaken for the Above

In many instances, beliefs that are thought to be special intuitions and the like might turn out to be mere IGBs. For example, some might consider a belief in other minds to be a special kind of immediate intuition, when (in a given instance at least) it might be just an IGB that results from an inexplicit reasoning process operating on observances of bodily behavior. Likewise, as W.T. Stace observes, belief in the independent existence of objects might not be an intuition but rather an IGB, one influenced by our propensity to choose simple and continuous interpretations of our surroundings (1945b, 140). One can also imagine a purported case of clairvoyance actually being an IGB. Perhaps Wednesday you heard on the radio that the president would be in New York on Sunday. By Sunday you had forgotten about the radio report, but you found yourself with a strong impression that the President was in New York. When later you read in the newspaper that the President was in New York, you mistakenly attribute your belief to clairvoyance.

D. Cooperation with other sources

While IGBs are distinct from the kinds of belief canvassed above, in any given instance, any of them *could* be among the inputs to the inexplicit reasoning processes that issue in IGBs. For instance, I might have an intuition from a moral faculty that it is always wrong to kill an innocent human being. Additionally, I might have other beliefs concerning the nature of conception and embryonic development that lead to an implicit belief (one I’ve never explicitly considered before) that an embryo is a human being. Then someone asks me if I think it is right to destroy unused embryos. An IGB is formed, and I respond, “No, that doesn’t seem right.” This IGB is the result of an implicit inference from

wearing a blue shirt with sleeves that only go so far down the arm, with horn-rimmed glasses, etc., etc. On the other hand, take away my IGB that Lydia is a liar, and it might be that no other IGBs arise from this experience.

5. Occurrent IGBs are often accompanied by some sort of sensuous imagery (due to the fact that concepts themselves tend to be accompanied with some sort of imagery).

a number of implicit premises, one of which derived from a faculty of moral intuition. I will not explore the existence of disputed faculties such as extra-sensory perception, or faculties of mystical, moral, or metaphysical intuition. However, if there are such faculties they may well provide input into the processes that result in implicitly grounded beliefs. For the purposes of this paper, however, we need assume no disputed sources of IGBs.

E. Thesis and comments on justification

An implicitly grounded belief is always perspectively basic—the one who holds it is neither able to reproduce the grounds for the belief, nor to manufacture a strong justifying argument for it. Many epistemologists would say one does not need argumentative support in order for one’s basic perceptual, memory, introspective, and *a priori* beliefs to count as warranted. One need not have argumentative support to know that it is sunny outside the window, that one had eggs for breakfast, that one has a headache, or that if all A are B, and all B are C, then all A are C. Need one have good reasons to know that a stock is ready to move, that it is best not to give money to a particular person, that a job applicant would not be an asset to the company, that a friend is angry, or that there is something wrong with one’s child? I argue in this paper that the answer is “no.” Some IGBs are *properly* perspectively basic, and of those that are basic, some are *properly* basic. IGBs can be justified and warranted. While it is impractical to survey all plausible accounts of justification and warrant to show that IGBs fulfill the criteria of each, below I argue that IGBs enjoy the main qualities required of justified and warranted beliefs by externalist, internalist, and aretaic accounts of warrant. I then discuss a few possible objections to this position. I now turn to a few brief comments about the concept of “justification.”

Some epistemologists define “justification” in an *internal* sense, as being within one’s epistemic rights or as believing well given one’s perspective. Internal justification does not fill the gap between true belief and knowledge. One’s beliefs can be true, and one can be rightly convinced that they are, and yet still fail to have knowledge because of factors not perceivable from one’s perspective, such as having an unrecognized cognitive defect, lacking crucial information, or being in misleading circumstances.

Other epistemologists use “justification” as that which *does* fill the gap between true belief and knowledge. Such justification involves not only believing well given one’s perspective, but something beyond one’s perspective falling into place as well, such as one’s belief being indefeasible, being formed reliably, or by being formed by properly functioning cognitive faculties. For the purpose of clarity, I will use “warrant” to refer to

internal/external justification, that which together with true belief yields knowledge, and I will use “justification” to refer to internal justification.

II. Truth-Conduciveness and Externalism

While some internalist theories hold that epistemic justification or warrant has nothing to do with truth, most theories, both internalist and externalist, hold that a justified or warranted belief is likely true.⁶ This might seem to be a problem for my thesis, since many people overestimate their intuitive abilities and form IGBs unreliably. For instance, Gilovich et al. studied the “hot hand phenomenon” on six basketball teams including the Philadelphia 76’ers and discovered that while players estimated that they shot better after a series of made shots than after a series of misses (and 9 out of 10 fans agreed), the facts were that players were actually slightly more likely to miss a shot after a series of successful shots than after a series of misses (Gilovich, Vallone, and Tversky 1985). Myers (2002) documents many people often are poor at forming what I have called IGBs, including intuitions about our past and future, about our own expertise, about social situations, about finances, about sports, and about clinical diagnoses. IGB-formation can be adversely influenced by factors such as hindsight bias, self-serving bias, loss aversion, the sunk cost effect, and confirmation bias to name a just a few.

None of these facts, however, impugn my thesis that many of our IGBs are justified and warranted. While there are many cases of poorly formed IGBs there are also plenty of cases of well-formed ones. Documented cases of reliably formed IGBs include the ability of most people to make accurate interpersonal judgements. Myers suggests that being able to quickly and accurately assess another person has had evolutionary survival value, and so it is a “small wonder that the first ten seconds of a relationship tell us a great deal, or that our capacity for reading nonverbal cues crosses cultures” (2002, 33). People will vary significantly in their intuitive abilities because of different cognitive abilities and styles. Likewise, various circumstances and kinds of preparation seem to enhance intuitive ability, such as an openness to experience, general study and immersion in a field, psychological freedom to explore, allowing for a period of non-intentional ‘incubation,’ and affective motivation to arrive at the truth (Monsay 1998, 116–117). For many of us, IGBs about matters with which we have experience such as what another person is

6. “According to this traditional conception of ‘internal’ epistemic justification, there is no logical connection between epistemic justification and the truth” (Chisholm 1988, 286). However, as Kihyeon Kim (1993) points out, a number of internalists, such as Lehrer and Bonjour, view the truth connection as a necessary component of epistemic justification.

thinking, about what a spouse is feeling, or whether a child is well are reliable. And in other, more remarkable and well-prepared people, the range of reliable IGBs is even more extensive.

Though the processes that produce IGBs *can* be unreliable in some cases, this is also true of those belief-forming processes that are universally acknowledged to be capable of producing justified and warranted beliefs—reasoning, perception, memory, and introspection. Since we do not withhold positive epistemic status from beliefs formed by such processes when they are in fact reliably formed, the unreliability of IGBs in some instances does not mean that IGBs cannot be warranted. Some people form IGBs more accurately than others, but in many cases processes that lead to IGBs are reliable, and thus IGBs can enjoy externalist warrant.

Consider one kind of externalism, defeasibility theory. As defeasibility theorists such as David Annis, Peter Klein, Keith Lehrer, and Marshall Swain indicate, the warrant for a belief can be compromised by contrary evidence that the subject does not have (See Annis 1973, Klein 1971, Lehrer 1990, and Swain 1981). For instance, Annis writes that a belief *h* is known when *h* is believed, is true, and there is “a set of statements *A* that fully justifies *S* in believing that *h* and there is no statement that defeats this justification” (Annis 1973, 199). Of the contrary evidence that might defeat a belief, some will be such that the subject might have access to it, but only through an IGB-formation process. For instance, I may not be able to prove my friend is upset through argument, though my strong gut feeling tells me that he is. This feeling may arise from facts about present and previous observations that I am not able to articulate. If I ignored my strong gut feeling and concluded that nothing is wrong because none of the propositional evidence I could muster proves otherwise, I would adopt an epistemic practice which in this and similar cases leads to unwarranted belief. Defeasibility theories suggest that in many cases the more evidence a subject can access, the less likely it is that her belief will be unwarranted. Thus, sometimes testing conscious reasoning against a source that takes more evidence into consideration will make sense if one desires to believe in a way that promotes an epistemically desirable set of beliefs.

Most contemporary externalists do add at least one *internal* constraint on warrant—a no-defeater clause. According to the no-defeater clause, for a belief to be warranted the subject must not believe (nor believe upon reflection) that her belief is defeated. Other versions substitute “justifiably believe” or “warrantedly believe” for “believe” in this definition. One might think that those who hold IGBs do have a defeater for their beliefs, namely the fact that the beliefs are perspectively basic. In the next section, I will argue

that the fact that an IGB is perspectively basic does not keep it from being justified or warranted.

The externalist emphasizes reliability—things going right from an external perspective—more than the subject herself explicitly understanding that things are going right. Given this perspective, IGBs arrived at in a reliable manner can be warranted. But are things more challenging for IGBs from an internalist perspective, and more particularly in light of the main motivation for internalism, deontology?

III. Internalism

A. Deontology

Deontological theories maintain that justification or warrant requires fulfilling one's epistemic duty or duties. Hilary Kornblith argues that one is obliged to responsibly seek truth and gather evidence. Chisholm holds that one has an epistemic duty to try to believe truths. In particular, he defines justification ultimately in terms of epistemic reasonability, and states that "epistemic reasonability could be understood in terms of the general requirement to try to have the largest possible set of logically independent beliefs that is such that the true beliefs outnumber the false beliefs" (Chisholm 1980, 7).

Chisholm and Kornblith typify most proponents of epistemic duty in that they enjoy believing in such a way, or preparing oneself to believe in such a way, as to attain the "epistemic truth goal." The epistemic truth goal is not merely to believe a large number of truths and a small number of falsehoods, for this would be best met by believing only simple mathematical truths, and avoiding all other thoughts. The goal of the epistemic life is to arrive at a sufficiently complex, comprehensive, and important set of true beliefs. IGB-formation can play a significant role in pursuing such a comprehensive, epistemically desirable set of true beliefs, for it can enable a person to form interesting, diverse, and important true beliefs that would not be arrived at by other means.⁷ I may know a friend is angry even when the evidence is so subtle that I cannot convince others of his anger. It may be that no matter how detailed my descriptions, my interlocutor may still favor another conclusion, perhaps that my friend has indigestion. I will be reduced to saying,

7. Eubanks et al. have discovered that intuition is uniquely capable of providing creative, original solutions to difficult problems (Eubanks, Murphy and Mumford 2010). Lewicki et al. (1992, 799) conclude from their research that "Our nonconscious information-processing system appears to be incomparably more able to process formally complex knowledge structures, faster and 'smarter' overall than our ability to think and identify meanings of stimuli in a consciously controlled manner."

“Look, I can’t explain it, but I could just *tell*. He *was* angry.” In such cases, propositions are inadequate for capturing all of the grounds for my beliefs. In cases of intuitive insight, we are able to expand our set of true beliefs in a way that would be impossible if we attempted to withhold our belief. Each such expansion lays the groundwork for acquiring more true beliefs that are related to the first. There are times, then, when forming and maintaining an IGB is more conducive to the epistemic truth goal than believing only what one can establish through argument.

One might think that even though forming an IGB can sometimes be a good way of reaching the epistemic truth goal, IGBs should still not be considered warranted because there is always a *better* alternative to forming an IGB. On this view, a belief is not warranted if one can expect to do better with respect to the truth goal by using a different belief-forming method. Many perceptual beliefs, then, are thought to be properly basic while IGBs are not because a good alternative to forming perceptual beliefs does not exist. Memory, rational intuition, and introspection are not well suited for producing perceptual beliefs. One might choose to reason discursively to a perceptual belief, but the result would not be epistemically superior to grounding the belief in perceptual experience. One might reason like this: I am having a perceptual experience with features *q*, *r*, and *s*. In past experiences, such features were truly indicative of a squirrel running up a tree. Therefore, I conclude that there is a squirrel running up a tree.

Notice three facts about such an odd technique of belief-formation. First, perception is still involved in this discursive process, and so if the purpose of arriving at the belief in this way is to avoid using perception that purpose is never realized. Second, the result of the discursive process is far less compelling than if the perception alone were to ground the belief. Finally, it certainly would not promote the epistemic truth goal for me to reach all of my perceptual beliefs through such a reasoning process even if I could. It would take so much time and effort, and often be so fruitless, that I would be able to expand my set of true beliefs far less while not improving its truth to falsehood ratio at all.

While superior alternatives to IGB-formation processes sometimes exist, in some cases, for some subjects, IGB-formation processes share with perceptual processes the three characteristics mentioned above. In these cases, any alternative belief-forming process will still involve an IGB, will be less compelling than if one’s belief were grounded solely in the IGB, and will be a poorer way of trying to achieve epistemic excellence. A person who is very intuitive in interpersonal matters may reliably form the strong IGB that a colleague is deceptive, and yet may be unable to specify the grounds for this belief. He could try to construct an argument for this proposition, but any available argument might be far less compelling than his clear and forceful IGB, and at any rate would be to

some degree dependent upon other IGBs, those which at some level would be required to support the implied premises for this argument. Likewise, the effort of trying to produce such an argument would require significant time and energy which, if the person had followed the intuitive path, might already have been used to produce a number of other interesting and important true beliefs. For this subject in this circumstance, the truth goal is best pursued by forming the IGB.

Since forming and maintaining an IGB can be a good way of pursuing the truth goal, forming and maintaining an IGB can be a good way of fulfilling epistemic duty, as may becoming more sensitive to and developing one's intuitive modes of thinking. If an intuitive interviewer is convinced that a job candidate would be a bad hire after talking with him, she may believe responsibly even if she cannot provide a justifying argument for this conviction. A genius at solving crimes may responsibly believe that a crime happened in a certain way, just because it seems strongly to him that it is so, even if he can give no good argument for this conclusion. Scott Johnston may believe responsibly when he concludes that a certain company's stock is about to move, even if this belief stems from what he calls a "sixth sense." Each of these individuals may be acting in the way that, given their abilities, experiences, and circumstances, best promotes the epistemic truth goal.

Some proposed epistemic duties go beyond requiring one to pursue the epistemic truth goal to specifying that one must do so in a particular fashion. For instance, Laurence Bonjour believes that we are obliged to "reflect critically on our beliefs" (Bonjour 1980, 63) and to accept "all and only beliefs which one has a good reason to think are true" (Bonjour 1986, 101). It should be clear from the preceding that Bonjour's requirement is biased in favor of a certain kind of thinker. He ignores the fact that some more intuitive people are able to believe responsibly even apart from critical reflection.

To claim that beliefs like "Barstow was poisoned" or "This stock is about to move" must be supported by good reasons in order to be warranted is to ignore differences among individual cognitive styles. Some people are clearly very accurate in believing in unconventional, more intuitive, ways. In some cases, these same people may be among the large portion of the human population that is only semi-skilled at conscious reasoning processes.⁸ The specific way in which an epistemic duty must be fulfilled, then, differs

8. For instance, studies have shown that in some cases elderly people are more accurate when they follow emotional and intuitive decision-making processes rather than explicit, reflective ones (Bruine de Bruin, Parker and Fischhoff 2012; Mikels et al. 2010).

from person to person. Epistemic duty must be understood in a way that is flexible enough to apply to all cases of responsible belief, intuitive as well as explicitly reflective.

B. Internalism and Special Access

Internalism requires a believer to have special access to the grounds of her belief, as well as perhaps to the adequacy of those grounds, in order for her belief to be justified.⁹ Since IGBs are perspectively basic by definition, an internalist might argue that IGBs cannot be justified. And since justification is thought by the internalist to be necessary for warrant, IGBs cannot be warranted either. I will argue that one who holds an implicitly grounded belief *can* have access to her grounds and their adequacy. This access is implicit, but as it turns out, we will see that there is no good reason for an internalist to claim that only *explicit* access will do.

For most internalists, the accessibility requirement seems to follow quite naturally from a commitment to deontology. Bonjour reflects this motivation for an access requirement: "One must accept all and only beliefs which one has good reason to think are true. To accept a belief in the absence of such a reason...is to neglect the pursuit of truth; such acceptance is, one might say, *epistemically irresponsible*" (Bonjour 1986, 101). To believe responsibly one must have a good reason, and this implies some kind of explicit access to the grounds for one's beliefs as well as to their adequacy.

Alston and Plantinga reject deontology and with it this explicit accessibility requirement (Alston 1988a; Plantinga 1993). If they are right to do so, then there is no challenge here to IGBs being warranted. But even if one agrees with a deontological conception of warrant, one need not insist that access to one's grounds and their adequacy be explicit. One can be a deontologist and still hold that IGBs can be warranted. In the previous section, I argued that even though an IGB is perspectively basic, one can still hold an IGB responsibly. I suggested, for instance, that Scott Johnston's belief that a stock is about to move might be responsible (and thus justified) even if he cannot provide a complete argument for his belief. How could this be so? Well, if Johnston is very intuitive, if he has a sixth sense, he will do better with respect to the truth goal by following his gut instincts. But, the deontologist will reply, is it enough for him to merely *have* this sixth sense if he is unaware of it? To believe responsibly, it would seem that he must have a good reason to think that he has this special skill and that the belief produced by it is well-grounded. Perhaps. But in what sense must he *have* this good reason? Need he be

9. For a treatment of the varieties of internalism, see Alston 1988a and Alston 1988b.

explicitly aware of it? Must he be able to become explicitly aware of it upon reflection? Neither is required in order for an intuitive person to believe responsibly. The intuitive person, naturally enough, will have an intuitive or implicit grasp of the well-groundedness of his IGB. He forms and maintains his IGBs when he has this implicit grasp of their well-groundedness and when such an implicit assurance is lacking, he at least tries to withhold or disbelieve, to do further research, to test hypotheses, etc. (Not only is the ability to make grounds explicit not necessary for warrant, in many cases it would be inadvisable, since many *intuitive* thinkers will arrive at a worse result through explicit reflection.¹⁰) Of course, if implicit assurance of the well-groundedness of one's beliefs comes in degrees, responsible believing may require various courses of action based on the degree (or even the type, if there are types) of intuitive assurance the believer has. Thus, one can plausibly be a deontologist and still hold that some implicitly grounded beliefs are justified and warranted. To do so would be to embrace an accessibility requirement, but to allow that one's access to grounds and/or their adequacy may be implicit in some cases.

The explicit access required by the internalist usually involves being able to reproduce the grounds of one's belief if asked. But why should the *accessibility* of grounds, as opposed to their being actually accessed, be an element in a belief's being justified or warranted? What good does a counterfactual do for the believer in her current believing? Isn't the important matter that her belief *be* well-grounded? If it is not, what does it matter that later she is able to produce grounds for that belief? If the believer produces an argument for a belief when questioned about it, she may be creating an argument for the belief that was never operational in the formation or maintenance of the belief before.¹¹ If that argument is a good one, then the belief might gain a justification that it did not have before, but there is nothing about *accessibility* to a good argument that would make an act of believing justified or warranted.

What is important is not *accessibility* of grounds and their adequacy, which could never make a belief justified or warranted, but some sort of functional *access* of those grounds by the believer *while* she believes. But to require in addition to access, that one must be able to produce grounds for the belief is too strong. If an intuitive person consistently forms true IGBs in a certain subject area, this suggests that she does have adequate access to the grounds of her beliefs, even if this access is implicit. Further,

10. For instance, Wilson and Schooler (1991) showed that intuitive self-knowledge for some people is compromised by explicit reflection.

11. Nisbett and Wilson (1977) and Wilson and Schooler (1991) argue that this is the case with regards to intuitive self-knowledge. Haidt (2001) argues that this is typically the case with regards to moral intuitions.

young children can have justified and warranted beliefs even when they are not able to articulate the grounds of their beliefs. In the end, we must acknowledge that one can have an implicit grasp of the well-groundedness of one's beliefs. This implicit special access is psychologically internal to the believer, and is really the only kind of access which can rightly be required for warranted belief.

Why has explicit accessibility been mistakenly thought to be necessary for justification and warrant? Perhaps because it does serve as a sign that one has the actual access to one's grounds and their adequacy which indeed *is* necessary for justification and warrant. If a person is able to provide a good argument for her belief if asked, then it makes sense to think that her belief well-grounded. I admit that explicit accessibility is often a decent sign of well-groundedness, but it is not a fail proof sign. It is not *sufficient* for justification, because a person might concoct an argument "on the spot" that was never involved in the production or maintenance of her belief. It is not *necessary* because there is no necessary connection between a beliefs being well-grounded (or one having an implicit assurance of this fact) and being able to explicitly produce grounds. The conviction that there is such a necessary connection seems to be the result of bias in favor of a certain kind of thinker. If explicit accessibility is not a legitimate necessary condition for justified belief, then internalism has no grounds for denying that implicitly grounded beliefs can be justified.

IV. Aretism and Proper Function

Virtue theorists such as Alvin Goldman, Ernest Sosa, and Linda Zagzebski believe that a warranted belief must arise from a truth-conducive virtue. As Goldman indicates, "Beliefs acquired (or retained) through a chain of 'virtuous' psychological processes qualify as justified; those acquired partly by cognitive 'vices' are derogated as unjustified" (Goldman 1992, 157). Goldman and Zagzebski offer informal lists of intellectual virtues and vices. Goldman's list follows:

I shall assume that the virtues include belief formation based on sight, hearing, memory, reasoning in certain 'approved' ways, and so forth. The vices include intellectual processes like forming beliefs by guesswork, wishful thinking, and ignoring contrary evidence. (Goldman 1992, 158)

IGBs appear neither on Goldman's list of virtues nor on the list of vices. (As was discussed earlier, IGBs are not the result of "guesswork."¹²) Given, however, that Goldman

12. Neither do IGBs involve the vice that Goldman lists, "ignoring contrary evidence." It is helpful, though, to distinguish between ignoring contrary evidence and believing *in spite of* contrary evidence. I may persist in

states that “belief-forming processes...are deemed virtuous because they (are deemed to) produce a high ratio of true beliefs” (Goldman 1992), it would seem that some IGB-formation processes should be considered virtuous by him. Zagzebski’s list of intellectual virtues includes qualities that are involved in the production of IGBs, such as the “adaptability of the intellect,” “being able to recognize reliable authority,” and “insight into persons, problems, and theories” (Zagzebski 1996, 114).

The *definitions* of intellectual virtue that these virtue theorists offer also suggest that IGBs can result from intellectual virtue, and so can fulfill the aretaic condition for warrant. Goldman describes intellectual virtue as a reliable process, where “process” is “construed as the sort of entity depicted by familiar flow charts of cognitive activity. This sort of diagram depicts a sequence of operations (or sets of parallel operations), ultimately culminating in a belief-like output” (Goldman 1992, 165). Zagzebski defines a virtue as “a deep and enduring acquired excellence of a person, involving a characteristic motivation to produce a certain desired end and reliable success in bringing about that end” (Zagzebski 1996, 137). Intellectual virtues for Sosa are “powers or abilities to distinguish the true from the false in a certain subject field, to attain truth and avoid error in that field” (Sosa 1991, 236). IGBs do result from processes such as Goldman describes, these processes are an entrenched and reliable acquired excellence as Zagzebski requires, and they can, as Sosa suggests, distinguish the true from the false in a particular subject field. The main distinguishing mark of an intellectual virtue on each account is that it leads reliably to true belief. The second aspect of intellectual virtue that each suggests is that it involves a settled disposition to form beliefs in a certain way. Since IGBs can fulfill these conditions, it seems that IGBs can in principle fulfill aretaic accounts of warrant.

Plantinga’s proper functionalism holds that a warranted belief must be formed in accordance with the subject’s cognitive design plan. The widespread and successful formation of IGBs in both common and academic life suggests that if there is a human cognitive design plan, IGB-formation is part of it. Plantinga acknowledges that people can differ significantly in cognitive equipment and methods. He is inclined to think that the Twins described by Oliver Sacks, who were severely mentally challenged and yet could perform amazing mathematical calculations with incredible rapidity, had knowledge of their conclusions (Plantinga 1993), and so, apparently, he would agree that different

believing that I was out of town yesterday even though witnesses say that they saw me at a local store. The contrary evidence in this case is defeated by my clear memory of being out of town, and so in this case I retain my memory belief in the face of contrary evidence. Such believing in the face of contrary evidence does not exhibit a vice but a virtue, and sometimes one who holds an IGB will exhibit this virtue.

subjects may have or acquire cognitive design plans, successfully aimed at truth, that differ from those had by the majority of people. A detective who intuitively arrives at amazing conclusions about crimes could be one such person, and Scott Johnston another. Their beliefs could be warranted even though many others would not be warranted in believing as they do given the same evidence.

V. Objection - The Slippery Slope

If we accept that IGBs can be warranted, must we say the same about all manner of odd beliefs, such as beliefs putatively formed by clairvoyance, telepathy, ESP, or past-life memories? (I will call these “paranormal beliefs.”) Aren’t IGBs and paranormal beliefs epistemically indistinguishable, and since paranormal beliefs should not be considered warranted, doesn’t this mean the same is true of IGBs? The answer to these questions is that IGBs *can* be epistemically distinguished from paranormal beliefs.

The main difference between IGBs and paranormal beliefs is that IGBs come about through a natural process, the contours of which are basically understood by cognitive psychologists. IGBs come about through the mind’s operation via induction, deduction, chunking, and the association of ideas on information gained over time from experiences mediated by normal human faculties—perception, memory, introspection, reasoning, and testimony. Beliefs purportedly formed through acts of clairvoyance, telepathy, ESP, and memories of previous lives have not yet been shown to have such a scientifically understandable natural basis. Thus, from an epistemological perspective, we have reason to doubt that such beliefs issue from reliable, stable dispositions of the subject, and are rooted in grounds that are in fact truth-conducive.

My basic IGB that Lydia is lying may well be grounded in perceptual cues that in the past have proven deceptive in people I’ve met. But what could possibly be the truth-conducive grounds of the belief that I used to be an Egyptian prince in a previous life? What memory traces in my brain or mind could give rise to such a belief? Given our current understanding of memory, there is nothing in which such a belief could be reliably grounded, and so such a belief seems unwarranted. The same can be said concerning the clairvoyant belief that right now Fidel Castro is listening to Miles Davis. If I have no information about Castro, about his habits, and about what he is doing now, through perception, testimony, or any other means, in what could the belief that he is listening to Miles Davis be grounded? Given what we know about human cognitive processes, no truth-conducive grounds exist for this belief. That IGBs come about through a natural,

scientifically understandable process is a reason to countenance their being warranted while still doubting that paranormal beliefs are warranted.

Sometimes paranormal beliefs are claimed to have “supernatural origins.” If so, such reputed mechanisms may have to be established by markedly different lines of inquiry. For our purposes at any rate, this would merely serve as another distinguishing characteristic between IGBs and paranormal beliefs. Finally, if it were scientifically verified that some people consistently form paranormal beliefs reliably, then even without a naturalistic explanation one might have to admit they were warranted. But in that case, any association with paranormal beliefs would not impugn the warrant of IGBs.

VI. Conclusion

Implicitly grounded beliefs are common, and especially so for intuitive people. As Stace indicates:

We have them about all subjects, about the physical world, about ourselves, about each other, about morals, about our breakfasts, about our work, our play, in short about everything. The whole fabric of human thinking, from our most trivial thoughts to our most profound philosophical treatises, is shot through with them. (1945a, 35)

I have argued that IGBs can be justified and warranted. They can be properly perspectively basic, and those that are basic can be properly basic. To hold that IGBs cannot be justified and warranted is probably to assume falsely too much homogeneity in cognitive equipment, and to assume an unwarranted bias favoring explicit reasoning processes. Forming and maintaining IGBs, as well as developing one’s intuitive abilities, can be particularly good ways of striving towards the epistemic truth goal, and so IGBs can fulfill the most prominent condition on warrant from externalist theories of warrant. For this reason, IGBs can also be deontologically justified. IGBs also fare well on internalism more generally because they arise in subjects who do have an effective form of internal access to the grounds of their beliefs, albeit an implicit one. W.T. Stace was right to argue that in the philosophical community, one cannot expect to base one’s views on IGBs and have them accepted by those who do not share one’s perspective. One should do whatever one can to make the grounds for one’s beliefs explicit in such contexts. But in order for individuals to have knowledge, such an ability to make one’s grounds explicit is not required.

References

- Alston, W. 1988a. "An Internalist Externalism." *Synthese* 74 (3): 265–283.
- . 1988b. "Justification and Knowledge." In *Proceedings of the XVII World Congress of Philosophy, Volume 5*. Montreal: Editions Monmorcency.
- Annis, D. 1973. "Knowledge and Defeasibility." *Philosophical Studies* 24 (3): 199–203.
- Barrett, L., M. M. Tugade, and R. W. Engle. 2004. "Individual differences in working memory capacity and dual-process theories of the mind." *Psychological Bulletin* 130 (4): 553–573.
- Berne, E. 1962. "Intuition VI: The Psychodynamics of Intuition." *Psychiatric Quarterly* 36 (1–4): 294–300.
- Berne, E. 1949. "The Nature of Intuition." *Psychiatric Quarterly* 23 (2): 203–226.
- Bonjour, L. 1986. "Can Empirical Knowledge Have a Foundation?" In *Empirical Knowledge*, by P. Moser, 95–115. Totowa, New Jersey: Rowan and Littlefield.
- Bonjour, L. 1980. "Externalist Theories of Empirical Knowledge." In *Midwest Studies in Philosophy V, Studies in Epistemology*, by P. French, T. Uehling and H. Wettstein, 53–73. Minneapolis: University of Minnesota Press.
- Bruine de Bruin, W, A. M. Parker, and B. Fischhoff. 2012. "Explaining adult age differences in decision-making competence." *Journal of Behavioral Decision Making* 25 (4): 352–360.
- Bunge, M. 1962. *Intuition and Science*. Englewood Cliffs, N.J.: Prentice-Hall.
- Chassy, P., and F. Gobet. 2011. "A hypothesis about the biological basis of expert intuition." *Review of General Psychology* 15 (3): 198–212.
- Chisholm, R. 1980. "A Version of Foundationalism ." In *Midwest Studies in Philosophy V, Studies in Epistemology*, by P. French, T. Uehling and H. Wettstein, 3–32. Minneapolis: University of Minnesota Press.
- Chisholm, R. 1988. "The Indispensibility of Internal Justification." *Synthese* 74 (3): 285–296.
- Crutchfield, R. 1960. "Male Superiority in 'Intuitive' Problem Solving." *American Psychologist* 15 (7): 429.
- Eubanks, D. L., S. T. Murphy, and M. D. Mumford. 2010. "Intuition as an influence on creative problem-solving: The effects of intuition, positive affect, and training." *Creativity Research Journal* 22 (2): 170–184.
- Evans, J., and K. Frankish. 2009. *In two minds: Dual processes and beyond*. New York: Oxford University Press.
- Eyesenck, H. 1990. "Implicit and Explicit Memory." In *The Blackwell Dictionary of Cognitive Psychology*, by H. Eyesenck, 185–6. Oxford: Basil Blackwell Ltd..

- Gilovich, T., R. Vallone, and A. Tversky. 1985. "The Hot Hand In Basketball: On the Misperception of Random Sequences." *Cognitive Psychology* 17 (3): 295–314.
- Gobet, F., P.C.R. Lane, S. Cheng, P.C.H. Croker, G. Jones, I. Oliver, and J.M. Pine. 2001. "Chunking mechanisms in human learning." *Trends in Cognitive Sciences* 5 (6): 236–243.
- Goldman, A. 1992. "Epistemic Folkways and Scientific Epistemology." In *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, by A. Goldman, 155–175. Cambridge, MA: MIT Press.
- Gore, J, and E. Sadler-Smith. 2011. "Unpacking Intuition: A Process and Outcome Framework." *Review of General Psychology* 15 (4): 304–316.
- Graf, P., and M.E.J. Masson. 1993. *Implicit Memory: New Directions in Cognition, Development, and Neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Griffin, T., S. Schwartz, and K. Sofronof. 1998. "Implicit Processes in Medical Diagnosis." In *Implicit and Explicit Mental Processes*, by K. Kirsner, C. Spelman, M. Mayberry, A. Obrien-Malone, M. Anderson and C. MacLeod, 329–41. Mahwah, NJ: Lawrence Erlbaum.
- Hadamard, J. 1954. *An Essay on the Psychology of Invention in the Mathematical Field*. New York: Dover Publications.
- Haidt, J. 2001. "The emotional dog and its rational tail." *Psychological Review* 108 (4): 814–834.
- Kaufman, S.B. 2009. "Faith in intuition is associated with decreased latent inhibition in a sample of high-achieving adolescents." *Psychology of Aesthetics, Creativity, and the Arts* 3 (1): 28–34.
- Kim, K. 1993. "Internalism and Externalism in Epistemology." *American Philosophical Quarterly* 30 (4): 303–316.
- Kirsner, K., C. Spelman, M. Mayberry, A. Obrien-Malone, M. Anderson, and C. MacLeod. 1998. *Implicit and Explicit Mental Processes*. Mahwah, NJ: Lawrence Erlbaum.
- Klein, P. 1971. "A Proposed Definition of Propositional Knowledge." *The Journal of Philosophy* 68 (16): 471–482.
- Kunst-Wilson, W.R., and R.B. Zajonc. 1980. "Affective Discrimination of Stimuli that cannot be Recognized." *Science* 207 (4430): 557–558.
- Lehrer, K. 1990. *Theory of Knowledge*. London: Routledge.
- Lewicki, P., T. Hill, and M. Czyzewska. 1992. "Nonconscious acquisition of information." *American Psychologist* 47 (6): 796–801.
- Mach, E. 1896. "On the Part Played By Accident in Invention and Discovery." *Monist* 6 (2): 161–175.

- Mikels, J, and T. Gilovich. 2013. "The Dark Side of Intuition: Aging and Increases in Nonoptimal Intuitive Decisions." *Emotion* 13 (2): 189–195.
- Mikels, J.A., C.E. Lockenhoff, S.J. Maglio, M.K. Goldstein, A. Garber, and L.L. Carstensen. 2010. "Following your heart or your head: Focusing on emotions versus information differentially influences the decisions of younger and older adults." *Journal of Experimental Psychology: Applied* 16 (1): 87–95.
- Monsay, E.H. 1998. "Intuition in the Development of Scientific Theory and Practice." In *In Intuition, The Inside Story: Interdisciplinary Perspectives*, by R. Davis-Floyd and P. Arvidson, 103–120. New York: Routledge.
- Myers, D.G. 2002. *Intuition: Its Powers and Perils*. New Haven: Yale University Press.
- Nisbett, R. E., and T. D. Wilson. 1977. "Telling more than we can know; Verbal reports on mental processes." *Psychological Review* 84 (3): 231–259.
- Plantinga, A. 1993. "Why We Need Proper Function." *Nous* 27 (1): 66–82.
- Poincare, H. 1946. *The Foundations of Science*. Lancaster, PA.: The Science Press.
- Rouse, W.B., and N.M. Morris. 1986. "On Looking Into the Black Box: Prospects and Limits in the Search for Mental Models." *Psychological Bulletin* 100 (3): 349–363.
- Russell, B. 1956. "How I Write." In *Portraits from Memory and other Essays*, by B. Russell, 194–198. London: George Allen and Unwin.
- . 1948. *Human Knowledge, Its Scope and Limits*. London: Allen & Unwin.
- Schachter, D. 1987. "Implicit Memory: History and Current Status." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13 (3): 501–518.
- Sosa, E. 1991. *Knowledge in Perspective*. New York: Cambridge University Press.
- Speelman, C. 1998. "Implicit Expertise: Do We Expect Too Much From Our Experts." In *Implicit and Explicit Mental Processes*, by K. Kirsner, C. Speelman, M. Mayberry, A. Obrien-Malone, M. Anderson and C. MacLeod, 135–48. Mahwah, NJ: Lawrence Erlbaum.
- Stace, W.T. 1945a. "The Problem of Unreasoned Beliefs I." *Mind* 54 (214): 27–49.
- . 1945b. "The Problem of Unreasoned Beliefs II." *Mind* 54 (214): 122–147.
- Strack, F., and R. Deutsch. 2004. "Reflective and impulsive determinants of social behavior." *Personality and Social Psychology Review* 8 (3): 220–247.
- Swain, M. 1981. *Reasons and Knowledge*. Ithaca, NY: Cornell University Press.
- Szalita-Pemow, A. 1955. "The Intuitive Process and its Relation to Work with Schizophrenics." *Journal of the American Psychoanalytical Association* 3 (1): 7–18.
- Tannous, P. 1997. *Investment Guru*. New York: New York Institute of Finance.
- Thrash, T., and A. Elliot. 2003. "Inspiration as a Psychological Construct." *Journal of Personality and Social Psychology* 84 (4): 871–889.

- Wallas, G. 1926. *The Art of Thought*. New York: Harcourt, Brace, and Co..
- Westcott, M. 1968. *Toward a Contemporary Psychology of Intuition: A Historical, Theoretical, and Empirical Inquiry*. New York: Holt, Rinehart, and Winston.
- Wilson, T. D., and J. W. Schooler. 1991. "Thinking too much: Introspection can reduce the quality of preferences and decisions." *Journal of Personality and Social Psychology* 60 (2): 181–192.
- Zagzebski, L. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge : Cambridge University Press.

Journal of Cognition and Neuroethics

The Enigma Of Probability

Nick Ergodos

Biography

I have studied logic, mathematics, physics, and philosophy. My academic interests are in the cross section between mathematics, physics and philosophy. A future project is to investigate the relationship between the probability concept proposed here and the logic of Quantum Mechanics in more detail.

Acknowledgements

I want to thank the unnamed referee of the *Journal of Cognition and Neuroethics* for many helpful suggestions. I also want to thank Zea Miller for proofreading my text.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Ergodos, Nick. 2014. "The Enigma Of Probability." *Journal of Cognition and Neuroethics* 2 (1): 37–71.

The Enigma Of Probability

Nick Ergodos

I can stand brute force, but brute reason is quite unbearable. There is something unfair about its use. It is hitting below the intellect.

— Oscar Wilde

Abstract

Using “brute reason” I will show why there can be only one valid interpretation of probability. The valid interpretation turns out to be a further refinement of Popper’s Propensity interpretation of probability. Via some famous probability puzzles and new thought experiments I will show how all other interpretations of probability fail, in particular the Bayesian interpretations, while these puzzles do not present any difficulties for the interpretation proposed here. In addition, the new interpretation casts doubt on some concepts often taken as basic and unproblematic, like rationality, utility and expectation. This in turn has implications for decision theory, economic theory and the philosophy of physics.

Keywords

Bayesianism, decision theory, expectation, expected utility, expected value, fair game, measurement problem, probability interpretation, St Petersburg paradox, two-envelope problem

Historical Introduction

Today we have a whole zoo of probability interpretations. We have propensity interpretations, frequency interpretations, objective Bayesian interpretations, subjective Bayesian interpretations, logical interpretations, personalistic interpretations, classical interpretations, formalist interpretations and so on, almost without end. These interpretations claim that the ontology of probability is either physical, psychological, epistemic, logical or mathematical. Some of them view probability as objective, others as subjective. Some even go to the extreme and say that probability is merely an empty word that we are allowed to interpret in any way we want, as long as we do not violate the axioms of probability.

To understand why we have this wild bouquet of philosophical interpretations it is necessary to study history. The cause of the confusion is an old gambling problem called the St. Petersburg paradox. As this problem is still unsolved, it continues to infuse

confusion. The various ways that have been proposed to escape this problem have led to the scattered philosophical situation we have today for the probability concept.

It is easy to state the St. Petersburg problem. Let us say we play a very simple game where you toss an ordinary coin until heads comes up. If heads comes up in the first toss you will get one dollar from me. If it comes up at the second toss you will receive two dollars from me and so on. We double the amount for each tails-toss you manage to get before you get heads and the game ends. The question now is what you would be willing to pay to play this game. The smallest amount you can win is one dollar so the game should at least be worth one dollar. But exactly how much more than one dollar?

The classical answer to these questions is to calculate the expected value of the game and make sure you do not pay more than that. The problem with this approach is that the expected value of this game is infinite. This means that whatever I demand you to pay for the privilege to play this game you should accept the offer. This is because any amount, no matter how big, is smaller than infinity. But this advice is just crazy. No one in her right mind would pay even a modest sum for this game. Something must be seriously wrong here, but what? This is the St. Petersburg problem.

There has been three main ways to attack this problem (Dutka 1988; Jorland 1987).

- (a) The advice to pay infinitely much for the game is actually correct in theory but in practice it is not. If you are lucky you could win more money than I could possibly pay you, and even more money than all the money in the world. Likewise, there is a possibility that heads never comes up and we run out of time, because none of us have unlimited time at our disposal. As there are limited resources of time and money in the world the game as stated cannot actually be played in the real world. There is no need to modify the theory—we only have to keep in mind the actual circumstances when it is used. The theory in combination with the actual physical constraints at hand will produce a correct, finite, result. I will call this the Finite World argument.
- (b) The advice is mathematically correct but human beings do not value money linearly as the mathematical theory implicitly assumes. What we need is a new theory that complement the mathematical theory whenever humans and money is involved. I will call this the Human Value argument.
- (c) The advice is not correct and the mathematical theory therefore needs to be changed or re-interpreted. A re-interpreted theory usually does not give any advice for actions at all. At least not for single cases. The theory might give some

advices for action if a large number of repetitions of a game is considered. Very few attempts have been made to change the mathematical theory itself. I am only aware of one attempt, which we will study separately. I will therefore call this the Reinterpretation argument.

That a simple game like this can lead to such a big discussion was disturbing. And it got even more disturbing as the years, decades and centuries passed without no one being able to solve it. Early on people drew the conclusion that the concept of expected values does not seem to be as natural and unproblematic as was first assumed. The concept of probability, however, is confined to have a value between zero and one, and can never be infinite. This makes this concept immune from ending up in infinity-paradoxes like this. It must therefore be safer to have at the core of the theory. Consequently, soon after the discovery of the St. Petersburg problem the concept of expected value was replaced by the concept of probability as the central concept of the theory. The theory that up to now had been called many things, but never something including "probability," started to be called the Theory of Probability by everyone. The earlier focus in the theory on how to bet on different games of chance was gradually replaced by a focus on pure probability problems.

This was a smart move. However, the concept of probability and the concept of expected value are mathematically very close to each other. If one of these two concepts has philosophical problems connected to it, the other one will have philosophical problems as well. Not exactly the same, as we will see, but similar. The different proposals on how to get rid of the St. Petersburg problem is the direct cause of why we have the probability interpretations that we have, and why they are designed as they are. The Finite World and Human Value arguments are connected to the subjective, personalistic and Bayesian interpretations of probability. The Reinterpretation arguments led to the frequentist, propensity and formalist views of probability. Incidentally, in recent decades the Human Value argument has also played a crucial role in the development of economic theory, game theory, decision theory, and rational choice theory (Samuelson 1977).

The contemporary way to view the St. Petersburg paradox is that it is a very important historical problem that has led to a number of important theories. The fact that the problem is still unsolved does not seem to bother anyone anymore. In fact, the common understanding today is that the problem really is solved, only that it is not possible to say which solution is the correct one... We are told that we are free to pick any of the proposed solutions and let it be the solution of our choice. Notice that the three solution strategies above contradict each other. They cannot all be correct at the same

time. Being a simple mathematical problem this is really an odd situation. In no other area of mathematics are we free to choose the solution to a problem ourselves, and whichever solution in a set of mutually contradicting solutions we pick, we will have picked a correct one. Incidentally, the solution we pick also, to a large extent, determines the probability interpretation of our choice, and vice versa.

However, it is relatively easy to see that the Emperor is naked, i.e., that none of the proposed solutions is a valid solution. By doing this all the theories, concepts and probability interpretations that rely upon these false solutions will quickly have to find new and fresh justifications. Or else they will die.

The Finite World Argument

To ban everything that contains an unlimited number of entities from the realm of the possible, as this argument does, is both too drastic and too feeble at the same time. It is too drastic because if implemented universally all of mathematics as we know it would break down. Even the ancient Greek mathematicians knew that every finite entity could be expressed as an infinite sum of entities as well. For example, if I go from point A to point B this can be described as either a finite number of steps or as an infinite number of steps. A finite description would be to simply count the steps I need to take to go from A to B. This adds up to the total distance between A and B. An infinite description could be this: I first go half the distance to B. From there I go half the distance of what is left. From there half the distance from what is remaining, and so on. The entire walk is then described by the infinite series half the distance + one quarter of the distance + one eighth of the distance and so on *ad infinitum*. This infinite sum of course equals the full distance. The point is that whichever way my walk from A to B is described, the total sum must always be the same. Obviously, the real physical distance cannot be dependent on how it is described. However, if we obey the Finite World argument we need to say that some descriptions, the infinite ones, are unrealistic. For these a finite cap has to be imposed for the number of steps that actually can be performed in reality. No matter how big a number we choose as the finite cap, the capped sum of steps will never equal the full distance between A and B. So according to some descriptions I walk the full distance between A and B, but according to others I cannot make the full distance. The Finite World argument thus makes the distance between A and B dependent on how it is described, which is absurd. It is easy to see that this example is not isolated but can be multiplied to every area of mathematics. If the Finite World argument is taken seriously, mathematics as we know it would break down. This is drastic.

The feeble thing is that the Finite World argument does not solve the problem it was set out to solve. We can easily construct a situation where no actual physical entities are infinitely many—and yet the St. Petersburg paradox is still present. Imagine a situation where we have a gambling hall with a number of different games. Some of the games are classical lotteries of various types. Others are variants of the St. Petersburg game with different payoff functions and fees for playing them. A set of players is invited to the gambling hall to try their luck at the different games and lotteries. The contestants are given an equal large amount of playing chips that can be used to pay the fees for the games in the hall. Each contestant is free to play as much or as little she wants at each lottery or game. If she decide to not play anything at all, that is fine too. If she wins any of the lotteries or games she will get her reward in ‘winning chips’ that can only be collected; they cannot be used for playing. The goal for each contestant is to have as many chips at the end of the day as possible. All playing chips that might remain, together with all the winning chips each contestant have won, are counted. The contestant who has the largest total collection of chips will get a nice prize—a car, say.

In this situation the Finite World argument is useless. The only physical prize present is a car, which is not infinite. The chips can be multiplied indefinitely and do not need to be physical, so no cap can be imposed on any of the payoff functions for any of the St. Petersburg games. This means that the expected value for each of the St. Petersburg games is exactly the same, i.e., infinite. But, obviously, it is better to play a St. Petersburg game with payoff function 1,000 chips, 2,000 chips, 4,000 chips, 8,000 chips, and so on instead of the original game with payoff function 1 chip, 2 chips, 4 chips, 8 chips, and so on, assuming the fee for the games are the same. However, the theory does not make any distinction between these games at all. This means that you are still stuck with the original St. Petersburg problem and the Finite World argument is of no use at all. This is feeble.

We have now showed that the Finite World argument can never lead to a real solution to the St. Petersburg problem. All solutions in this category are thus false.

The Human Value Argument

The idea here is that ordinary humans do not think like mathematicians. In particular, ordinary men do not value large amounts of money in the way mathematicians do. For this reason ordinary humans do not have to obey the mathematical rules the theory produces. The theory of expected value is certainly correct mathematically, but it is simply not correct as a model for how ordinary men actually behave.

Ordinary men are driven by how happy they can get, not directly by how much money they can get. Sure, more money will probably make you more happy, but will twice the money you already have make you exactly twice as happy as you already are? Probably not. Twice as much money does not mean you will become twice as happy, and this psychological effect becomes even truer if your fortune is multiplied even more times. It seems to be some kind of law that the human appreciation of more money increases more slowly than what the actual amount of money possessed would indicate. A mathematical function that describes the decreasing additional happiness, or utility, you might get for increasing amounts of money is today called an utility curve.

If the utility, as described by a suitable utility curve, is inserted instead of the actual amounts of money you could win in the St. Petersburg game, the expected value, or rather the expected utility, becomes finite. This explains, according to this argument, why ordinary humans are not willing to pay what the mathematical theory demands in this case, but only a small finite amount.

This approach raises a number of new questions. Which mathematical function describes best the human utility of money? Should it be a universal curve for all men (except, perhaps, for mathematicians) or should it be different curves for different individuals? Can these curves be determined empirically? If so, how?

Interestingly, these questions have, over the years, not only been investigated thoroughly—the very idea of expected utility as a guide for human action has been hugely influential as a foundational concept. Almost all modern economic theory, decision theory and theory of rational choice rely on this concept. For example, in economic theory it has a prominent place in the “Law of diminishing marginal utility” and in decision theory in the “Expected utility hypothesis.” The concept of expected utility is also a vital part of the Bayesian interpretation of probability.

Despite its fame, the concept of expected utility never solved the St. Petersburg problem that it was originally designed to solve. This is easily seen by the following example. In the original St. Petersburg game, instead of winning one dollar, two dollars, and so on, let the payoff function be in “utility units” instead of in money. Instead of winning two dollars you win the amount of money that exactly makes you twice as happy as you would be if you won one dollar, and so on. The expected utility will in this case be infinite, and the utility argument is of no use anymore as all payoffs already are utilities. Hardcore believers in expected utility theory, when confronted with this example, usually resort to some arbitrary Finite World argument to escape the embarrassment. However, the very reason utility curves were invented in the first place was to construct a solution to the St. Petersburg problem that did not have to resort to silly Finite World arguments.

Others, critical to the concept of expected utility, try to solve the St. Petersburg problem using a concept of risk. It is because we are afraid to risk too much of our money that we are reluctant to pay a lot for playing the St. Petersburg game, even if the theory tells us that we should. This is another Human Value argument. Only our imagination sets the limits on how many different Human Value arguments we can invent to solve the St. Petersburg problem. However, there is a simple way to show that all Human Value arguments must be wrong, even those not invented yet. If it was the case that the St. Petersburg problem could be solved by a theory of human valuation of money, be it expressed in utilities or risk or whatever, it would be impossible to give an account of the St. Petersburg problem when neither humans nor money are involved. But this is indeed possible.

Consider a membrane which during one second transmits one hydrogen molecule with probability one half, two hydrogen molecules with probability one quarter, four hydrogen molecules with probability one eighth, and so on. How much hydrogen gas can we expect to be transmitted through the membrane during one second?

This is exactly the original St. Petersburg problem but now without any reference to neither humans nor money. This means that all solutions that rely on how humans value money are indeed useless in solving the St. Petersburg problem. This simple thought experiment shows that the Human Value argument can never lead to a real solution of the St. Petersburg problem. All solutions in this category are thus false.

The Reinterpretation Argument

Almost all aspects of the original theory have been reinterpreted just to avoid the St. Petersburg problem. The simplest one is merely a linguistic trick. Mathematicians do not talk about *infinite expectations* anymore. That way of talking is abandoned. What they say, instead, is simply that the expected value in that case *does not exist*. Something that does not exist cannot give you any advice, can it? Only expectations that do exist, i.e., is finite, can give you any advice. This trick resolves the St. Petersburg problem because the theory does not give you any advice at all as the expected value simply does not exist.

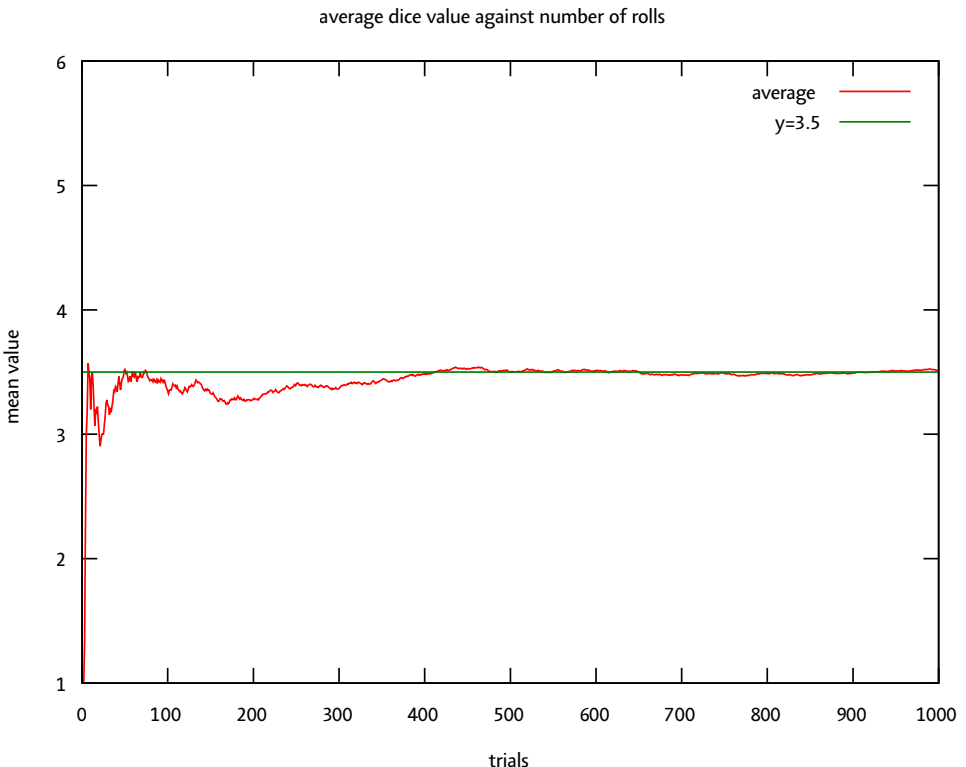
The problem with this escape route is that the distinction between when the theory will give you advice and not becomes arbitrary. This is because the difference between a situation where the expected value “exists” or “does not exist” can always be made arbitrary small. Consider for example the St. Petersburg game played with a real coin. It is an empirical fact that one of the sides of a real coin is always slightly more probable than the other side. The difference can be very small, but still there is always a difference. If we

Ergodos

happen to choose the side which is slightly less probable than $1/2$, as the side we repeat until the other side shows up and the game ends, we will end up with an expected value for the game that is finite and “exists.” If we happen to choose the other side as the one that we repeat, we will end up with an expected value that “does not exist.” Usually we do not know which side of a real coin that is the slightly more probable one—we just pick one of the sides at random as the one to repeat. So in half of the cases when we play this game the theory does have some advice to give us, while in the other half it does not. It is thus totally arbitrary when the theory can give us some advice and when it cannot.

To avoid this arbitrariness some reinterpret the theory even further so that the theory never gives any advice at all, whether or not the expected value exists. This does not lead to the same arbitrariness as before, but a theory that does not give any advice at all, in any case, is a little strange. Why do we have a theory in the first place if it does not have any practical applications?

Some other thinkers in this category say something really interesting. They claim that the expected value gets its entire meaning from the Law of Large Numbers, which is the name of the observation that the average gain will approach the expected value, or



mean, after a large number of repetitions of a game. The only thing we mean when we say that a game is fair is that in the long run we will win as much as we will lose. Say that we play the game where we get one dollar for each dot that shows up when throwing an ordinary die. The expected value for this game is 3.5 dollars. The Law of Large Numbers will guarantee that we will approximately break even in the long run when playing this game over and over for a 3.5 dollars fee. In the limit when we play an infinite number of games we will break even exactly. See the graph above.

Expected values so defined do not say anything about single cases. Expected values for single cases are viewed as unreliable or even meaningless. This makes the very question what one should pay for the St. Petersburg game a meaningless question. Only if we play the game over and over is it possible to know what we should pay. And indeed, if we play the St. Petersburg game an infinite number of times we should actually expect to win an infinite amount of money, exactly as the expected value shows. Adopting this statistical viewpoint resolves the paradox, according to this view.

This idea is very clever. The Law of Large Numbers will actually guarantee that the expected value reinterpreted in this way will always keep what it promises. A gambler paying the expected value for any game will with probability one gain as much money as she loses in fees when the number of games goes to infinity. This resolution of the St. Petersburg problem does not lead to inconsistencies or arbitrariness. But it still has major drawbacks why it cannot be viewed as a correct solution to the problem.

It totally misses the original question, which is to answer what a single game is worth. In particular, we still have no clue what to pay for the St. Petersburg game if it is offered only once. This reinterpreted expected value can only answer a restated St. Petersburg problem: What is the fair fee for each round if we play an infinite number of St. Petersburg games?

In addition, we have no clue how many times we need to play any given game in order to have permission to use the expected value as a fair fee. If this problem is avoided by saying that one should always play infinitely many rounds to have permission to view the expected value as a fair fee, well then suddenly all games imaginable are dealing with infinities. The problems originally attached to games with infinite, sorry "not existing," expected values are now affecting all games, also those that never caused any problems before. This is hardly an improvement of the situation.

Even if this solution of the St. Petersburg problem is the best so far, it is still very far from a correct solution. As mentioned at the beginning, there is also one additional proposal in this category that we need to treat separately as it is not merely a reinterpretation but actually a bold idea by William Feller to change the very definition

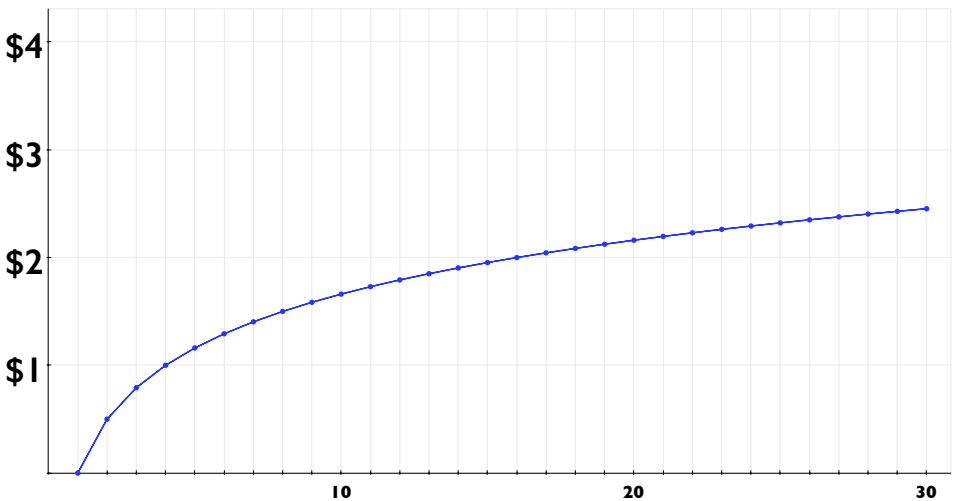
Ergodos

of expected values (Feller 1945; 1950). Even for him the Law of Large Numbers is what ultimately gives expected values their meaning. He slightly restates the mathematical expression for the Law of Large Numbers. His new expression is a true generalization of the classical Law of Large Numbers because, whenever the expected value is finite, his generalized expression becomes the ordinary expression. But whenever the ordinary expected value is infinite, or “does not exist,” his formula produces a series of variable fees. This idea is truly innovative. Never before did anyone call into question the implicit assumption that the fair fee for a game must be a constant.

According to Feller, the fair prize for the St. Petersburg game is

$$\frac{1}{2} \log t$$

where t is the number of times the game is played and \log is the logarithm to base two. See the graph below.



Let us say we decide to play the game 256 times. We should then be willing to pay 4 dollars per game according to Feller, as $\log 256$ is 8. If we plan to play 4096 times the fee we should be willing to accept is 6. By this we immediately see that when the number of games goes to infinity the fair prize also goes to infinity, which is exactly what the statistical reinterpretation of expected value says. But now, with Feller, we suddenly know a lot more what the game is worth even if we do not play infinitely many times.

This solution to the St. Petersburg problem is by far the best presented by anyone. But, unfortunately, it raises more questions than it answers.

First of all, is it his intention to replace the old concept of expected value with this new concept everywhere? It does not seem so. Instead, this concept seems to be tailor-made for games of the St. Petersburg type, i.e., for which the ordinary expected value is infinite, or “does not exist.” In that case his new concept falls into the same trap as the first reinterpretation idea we considered; it will be totally arbitrary when this concept and the old one should be used. The example with the St. Petersburg game using a real coin will apply equally well in this case.

Secondly, his concept is not applicable for one or even a few repetitions of a game. It is obviously not correct for a single game as $1/2 \log 1$ is 0 dollar, and we know by the construction of the game that it is at least worth one dollar, not nothing. This means that we know for sure that it is false for small t . He even admits this explicitly himself. But when, at what number t , is his formula beginning to be trustworthy? He does not say a thing about that. If we try to avoid this difficulty by simply say that t needs to be infinitely large for the formula to start to be trustworthy, then we have gained nothing by using a new concept that allows for variable fees.

Thirdly, we might ask ourselves why it should be, in any sense, *fair* to pay any of the finite fees suggested by his formula. It is, after all, only in the limit when t goes to infinity that his new concept obeys the Law of Large Numbers and the fees becomes “fair in the classical sense,” as Feller puts it. For no finite value of t is his concept fair in any sense. This means that even Feller falls into the same trap as the statistical reinterpretation; we have to play an infinite number of rounds in order to know that the suggested fee is a fair fee.

Despite his innovative and radical approach, we see that even Feller fails to solve the St. Petersburg problem. This concludes the task of showing that there still does not exist a single acceptable solution to the St. Petersburg problem. We can safely establish that the St. Petersburg problem is an open problem.

Probability Interpretations

Initially, it was a good move to replace the central concept of the theory—expected value—with the concept of probability. Probabilities seemed to be totally unproblematic, which was not exactly the case with expected values, as we have seen. However, after a while, it became evident that the concept of probability is problematic as well. Classically, the probability of an event is defined as the number of favorable cases divided by the total number of cases. For this to work we need to find atomic cases which are equally likely. For example, the probability of getting an even number when throwing an ordinary die is three over six, because there are three favorable cases (2, 4 or 6) and six equally

likely cases. But what does it really mean to be “equally likely”? It must mean that they are equally probable. But “probability” is the concept we try to define here and is of course forbidden to use before it is defined, or else the definition becomes circular.

To save the classical definition a new principle is introduced—the Principle of Insufficient Reason. It states that if you do not have sufficient reason to believe that one possible case is more likely than any other you are entitled to assign the same probability to each case. This formally removes the circularity in the definition of probability. However, this principle leads in itself to problems that are even worse—contradictions. Depending on the way we describe a situation, we will end up with different probability assignments for the same event. In addition, the principle cannot handle events where we have a continuum of outcomes. Even in these cases the principle leads to contradictions, as was first noted by Joseph Bertrand. We cannot have a principle at the core of our theory that leads to contradictions. It is evident that we need to give up this initial definition of probability entirely. It cannot be rescued. This is no good news at all. The concept of probability replaced the central concept of expectation just in order to give the theory a solid conceptual foundation. And now this solid core has evaporated into thin air, leaving a big conceptual hole at the heart of the theory. We are left with a mathematical theory that we have no clue what it is all about.

This situation needs to be solved in some way. But how? Very much as in the case with the St. Petersburg problem, a set of very different solutions to the problem emerges. In fact, it is because of the St. Petersburg problem people start to run in different directions when trying to find a solution to this problem. Depending on which strategy you settled for regarding the St. Petersburg problem, you will have different needs regarding the probability concept, and hence will end up advocating different definitions of probability.

If you believe in the Reinterpretation argument where expected values only are given a statistical interpretation as an average in a long run of games, you will end up with a frequentist interpretation of probability. According to this interpretation probabilities have no meaning for single cases, only a long—indefinitely long—sequence of events from repeatedly perform an experiment can be attributed a probability. The probability is then defined as the relative frequency with which the outcome you want to measure occur in the sequence.

Notice how close this definition of probability is to the corresponding proposed solution of the St. Petersburg problem. In both cases it demands that we repeat the event we are interested in infinitely many times, no matter what event it is. Both concepts get their ultimate meaning from the Law of Large Numbers. Another interesting thing to note is that this definition turns the classical relationship between probability theory and

statistics upside down. Probability theory is now ultimately based on statistics and not the other way around.

If you are of a more radical type and go for the non-interpretation alternative where expected values have no moral implication at all, you will end up having the same view on the probability concept as well. Anything that fulfills the standard axioms of probability will be an admissible probability to you.

If you believe in one of the Human Value arguments you will end up being a Bayesian or adhere to a logical or epistemic interpretation of probability. In these cases the concept of probability does not derive its meaning from the Law of Large Numbers, but from another theorem in probability theory—Bayes' theorem. As it stands, Bayes' theorem is totally uncontroversial. It only becomes controversial when placed as the foundation for every application of the concept of probability, even in cases when the prerequisites of the theorem are not fulfilled. The idea is to try to rescue the Principle of Insufficient Reason in a way that does not lead to contradictions. This is accomplished via an ingenious idea. The subjectivity of the utility concept is here transferred to the probability concept itself. You are thus free to assign any probability you want to any event, and it does not have to be in accordance with anyone else's assessment of the same event. Instead of trying to create a theory that in itself is free from contradictions, the burden of consistency is placed on each individual. It is up to you to make sure that none of your probability assignments lead to contradictions! This is where the extended version of Bayes' theorem enters the stage. If you feed the theorem with your initial beliefs, what the theorem produces will always keep you safe from causing any internal inconsistencies. Your probabilities will of course still contradict other person's probabilities, but that does not matter. The only important thing is that your internal sets of beliefs are not contradictory.

Note how this definition puts the relationship between probability theory and its users upside down. Instead of giving advice to its users on how to play games, the users now, so to speak, give advice to the theory or feed the theory. The individuals hold the ultimate truth themselves and the theory of probability merely functions as a set of traffic rules for how the individuals should think so that they do not end up having two different thoughts that collide. What happens if you do not follow the traffic rules is obvious—you become irrational. And who wants to be irrational? This idea of absolutely rational agents is the basis for rational choice theory as well as for most of modern economic theory. In a free market the agents are assumed to act in an absolutely rational manner, i.e., according to exactly the same concept of rationality as used in Bayesianism.

While the Principle of Insufficient Reason only could handle situations where all cases were equally likely, Bayesian theory does not have this limitation. Your personal estimates

of probabilities for different events can have any values whatsoever—they do not have to be the same for each case. Mathematically, this is expressed with a prior probability distribution, as the Bayesians denotes a subjective valuation function. This function can have any shape you like; it does not have to be a constant function with the same value everywhere. For example, if you have a die in front of you that you suspect is loaded, you will naturally assign probabilities to the sides that are not $1/6$ for each side.

This brings us to another basic principle of the Bayesian philosophy. If you suspect that the die in front of you is loaded, you have to take this into account when you set up your prior probability distribution. If you do not, you will easily contradict yourself. In fact, you have to include exactly everything you know at every instant when you make assessments regarding probabilities. It is not that easy to be a Bayesian.

There are many variants of Bayesian probability, where some are called logical, epistemic, or objective Bayesian probability. However, the basic idea is the same but emphasis on what is important is placed at different places in respective philosophy. Objective Bayesians, for example, believe that there are objective ways to construct the prior distribution function, which removes the subjectivity from the theory. Logical interpretations stress the idea that probability, with the help of Bayes theorem, should be viewed as an extension of ordinary logic, namely the logic extended to uncertain statements. Epistemic probability stresses the importance of evidence as the basic force behind probability assignments.

What they all have in common is a desire to be able to assign probabilities to also single cases, something the frequency and statistical interpretations fail to do. Attempts to extend the frequency approach to account for single cases as well are usually called Propensity interpretations of probability. Physicists have always been happy with the statistical concept of probability for their needs—until Quantum Mechanics entered the scene. Now, suddenly, a physical theory made probability statements about single events, which the old statistical concept of probability never can give an account for. Karl Popper started the movement of creating a propensity probability concept. One central goal is to extend the statistical theory of probability so that it becomes compatible with how probabilities are used within Quantum Mechanics.

Moving Forward

As we have seen, we have a strong historical and conceptual connection between the attempts to solve the St. Petersburg problem and all the current probability interpretations. We have also seen that the St. Petersburg problem is an open problem. But showing that

the St. Petersburg problem is still unsolved does not, by itself, prove that all probability interpretations are incorrect. It could be that one of the interpretations, while failing to solve the St. Petersburg problem, still has some justification as a probability interpretation. What we need is another example that explicitly shows how all current interpretations fail. Luckily, we have such an example. In recent decades another thought experiment has been discussed intensely, resulting in an even more confused debate than for the St. Petersburg problem. Also in this case, people are not able to agree upon what type of problem it is. Is it a problem within mathematics, logic, economics, psychology, or something else? Like the St. Petersburg problem it is a very easy problem to state, and yet no one has been able to solve it. However, the reason for this is very simple. None of the existing probability interpretations can be used to solve this problem, as we will see.

The Two-Envelope Problem

Imagine that you have two sealed envelopes in front of you containing money. One contains twice as much money as the other. You are free to choose one of the envelopes and receive the money it contains. You pick one of them at random, by tossing a coin, but before you open it you are given the option to switch to the other envelope instead. Do you want to switch?

Obviously, the situation is symmetric, so it cannot make any difference whatsoever if you stick or switch. And yet, there is a clever argument that shows that you should switch. Assume that the envelope you picked contains A dollars. The other envelope must either contain $2A$ or $A/2$ dollars depending on if A is the larger or smaller amount. You tossed a coin, so you know for sure that it is a 50/50 chance for either case.

There is a 50% chance that you get $2A$ and a 50% chance you get $A/2$ by switching. The expected value of switching is therefore $1/2 \times 2A + 1/2 \times A/2$ which is $5A/4$, or $1.25A$. This is more than A , what you already have, so you should indeed switch to the other envelope. But this clearly cannot be the case as the situation is symmetric. If you had selected the other envelope instead the same argument would have told you that you should have taken the envelope you now hold in your hand. The Two-Envelope problem is to spot and explain the flaw in this argument.

The frequentist or statistical concept of probability cannot even begin to solve this problem because it is evident from the setup that this is a single case. Propensity theories are of no help either. Those who think they can solve this problem are the Bayesians. The first Bayesian response is usually that the problem as stated is impossible to set up, as an infinite uniform prior distribution of money is implicitly required, and that is not

permissible according to Bayesian theory. Already this is a bit strange as many Bayesians have, in other contexts, argued that improper priors, as this is called, should indeed be allowed. To see why we end up with this improper prior, think about how the two envelopes could have been filled with money in the first place. If you pick an envelope that contains A and it is equally likely that the other envelope contains $2A$ or $A/2$ it must mean that there were initially, before you were offered to pick an envelope, two envelope pairs $\{A/2, A\}$ and $\{A, 2A\}$ where both of them were equally likely to be picked as the pair you got in front of you. Only in this case is it equally probable that the other envelope contains $2A$ as it is that it contains $A/2$. But this must be the case for each of the possible amounts in all envelopes, in each possible envelope pair, so there must have been an infinity of envelope pairs for the person setting up the game to choose from: ... , $\{A/8, A/4\}$, $\{A/4, A/2\}$, $\{A/2, A\}$, $\{A, 2A\}$, $\{2A, 4A\}$, $\{4A, 8A\}$, Each of the pairs must have had an equal probability to be chosen as the pair you got in front of you. This produces the improper distribution we talked about, as every envelope pair will have probability zero and yet when summing them all they must add up to one.

However, this solution can be escaped by changing the Two-Envelope problem slightly, by introducing an explicit prior probability distribution for envelope pairs that is not an improper distribution, but a proper one. Such a distribution cannot be a uniform distribution why adjacent envelope pairs will have slightly different probabilities. However, when carrying out the calculations we will still get the conclusion that we should switch to the other envelope whatever we find in the first envelope. But as we know that this will always happen we do not have to open the first envelope we pick. We already know in advance that the other envelope is slightly better. This is truly paradoxical. And yet, the situation is as symmetric as before so any calculation leading to a difference between the envelopes must be false. In this case the Bayesians cannot blame us for implicitly assuming improper prior distributions anymore. The prior probability distribution here is explicit and proper.

To escape this trap the Bayesians usually still blame it all on the prior distribution. But not for being improper but for having an expected value that is infinite, or an expected value that does not exist if you will. This is a quite strange argument. The question if improper priors should be allowed or not has been discussed among Bayesians as long as Bayesianism has existed, but now suddenly it is not possible to use probability distributions within Bayesian theory which lack expectation. There exist whole families of standard probability distributions that are used every day by statisticians all over the world that lack expected value and variance. The Cauchy distributions, for example, is one such family of distributions. Bayesians are usually proud of being able to extend the

application of probability from the narrow set of applications the statistical or frequentist interpretation can offer. Here we see an example of the opposite trend. Important probability distributions that are totally unproblematic for frequentists to use are banned by Bayesians.

Most Bayesians are nevertheless happy with this explanation of the Two-Envelope problem. They are confident that new even more evil versions of the paradox will not appear. This is because a theorem shows that in order to produce the paradox the prior distributions need to have an infinite expectation. But a version of the paradox was already published early on by Raymond Smullyan where we cannot blame a prior probability distribution for being the culprit (Smullyan 1992). In fact, this version does not need any probabilities at all, so no probability distributions whatsoever, prior or otherwise, enters the scene.

Consider the same setup as in the original problem where you are given the option to pick one of two envelopes where one contains twice as much as the other. The following two plainly logical arguments lead to conflicting conclusions:

1. Let the amount in the envelope you chose be A . Then by switching, if you gain you gain A but if you lose you lose $A/2$. So the amount you might gain is strictly greater than the amount you might lose.
2. Let the amounts in the envelopes be Y and $2Y$. Now by switching, if you gain you gain Y but if you lose you also lose Y . So the amount you might gain is equal to the amount you might lose.

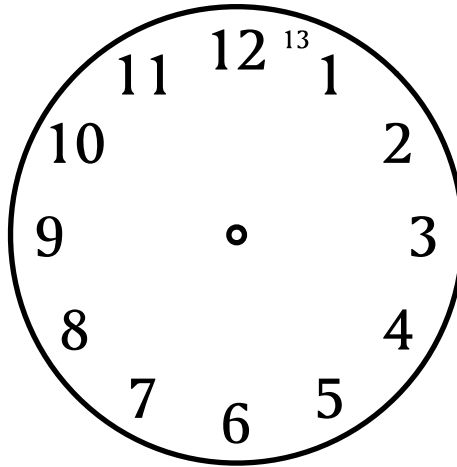
The usual Bayesian response to this version is that this is another problem than the original *because* it does not include probabilities. This is a strange argument. If we can preserve the paradox while removing one concept that we initially thought was vital from the account, we have indeed learnt something. In this case, the Two-Envelope paradox is not dependent on the concept of probability, at least not a Bayesian concept of probability.

Incidentally, it is indeed possible to construct a probabilistic variant of the Two-Envelope problem that neither include improper priors nor priors with infinite expectations.

Jailhouse Clock

Imagine that you find yourself in death row in a prison in Texas for a crime you did not commit. You do not know when you are going to be executed. On the wall there is a

clock that you have noticed is a bit odd. It works properly except when the small hand is between noon and 1 PM, and between midnight and 1 AM. What should take one hour here always takes exactly two hours. Apart from this the clock works as normal. This means that the clock needs 13 hours to complete a full cycle.



A prison guard enters your cell and tells you that the time for your execution and another fellow prisoner has been determined. He hands over two letters with the execution orders containing the time of execution. You are free to pick any of the letters and the one you pick will determine when you will be executed. You pick one that says that you will be executed at 4 o'clock some day, but the actual day is not specified. Then suddenly the prison guard has a big grin over his face. He says that he is in a good mood today and want to give you an offer. If you want you are free to take the other execution order instead. It is a really good offer because the probability is exactly $1/2$ that your time left in life will be doubled, and with probability $1/2$ that it will be cut in half.

The guard is ignorant of what is stated in the letters. He only knows the procedure for how to pick times for execution from the clock on the wall. Each hour has an equal chance of being selected for an execution. Execution orders are always created in pairs where one time left to live for a prisoner is twice as long as the other. This is possible to do via the clock on the wall in a cyclical manner, due to its odd feature. On that clock, twice of 1 is 2, twice of 2 is 4, twice of 4 is 8, twice of 8 is 3, twice of 3 is 6, twice of 6 is 12, twice of 12 is 11, twice of 11 is 9, twice of 9 is 5, twice of 5 is 10, twice of 10 is 7 and twice of 7 is 1. The pairs of letters presented to the prisoners are thus either {1, 2} or {2, 4} or {4, 8} or {8, 3} or {3, 6} or {6, 12} or {12, 11} or {11, 9} or {9, 5} or {5, 10} or {10,

7} or {7, 1}. Twelve different pairs are possible in total. The prior probability for each pair of letters to be selected is $1/12$. Because of this construction, when one letter in a pair is opened it is exactly as probable that the other envelope contains twice as much time left alive as it is half as much. In your case seeing the time stamp "4 o'clock" reveals that the only possible pairs of letters are {2, 4} and {4, 8}, and they are exactly equally likely to have been picked. The guard can therefore be absolutely certain that what he just told you is correct.

You are a Bayesian and know how probability theory can help you here. You have absolutely no reason to doubt that the other letter has either the time stamp "2 o'clock" or "8 o'clock" with probability $1/2$ each. By switching letter you will double your time left or cut it in half. What you can gain is thus twice of what you can lose, with an exact 50/50 chance for each outcome. Selecting the other letter will increase your chances of being released much more than it will decrease it. You know your lawyer needs every extra day she can get to reopen and win your case, before it is too late. So you would really need some extra time alive. On the other hand, of course, the situation is completely symmetric between the two letters you are given. No matter which one you would have opened first the other one would seem the more attractive. In particular, your fellow prisoner who got the other execution letter is also a Bayesian and reasons in the same way as you do. Both of you think it is a great opportunity to take the other letter so you end up swapping your execution letters.

None of the existing probability interpretations can solve this paradox.

Two Old Problems

In the early 1650's, four gentlemen went on a three-day trip from Paris to Poitou, in the mid-west of France (Todhunter 1865). During the trip they discussed interesting philosophical topics such as the ontological nature of infinity, the existence of the infinitely small, the nature of the number zero and the existence of absolutely empty space. Above all, they discussed the nature of mathematics in general and its connection to reality. It was during these discussions that one of the gentlemen, Antoine Gombaud, alias chevalier de Méré, brought up two gambling problems that he thought strengthened his philosophical position. His stance was that mathematics is a very beautiful art on its own but cannot, in general, be trusted when applied to the real world. In particular when mathematical reasoning tries to embrace the mysterious concept of infinity.

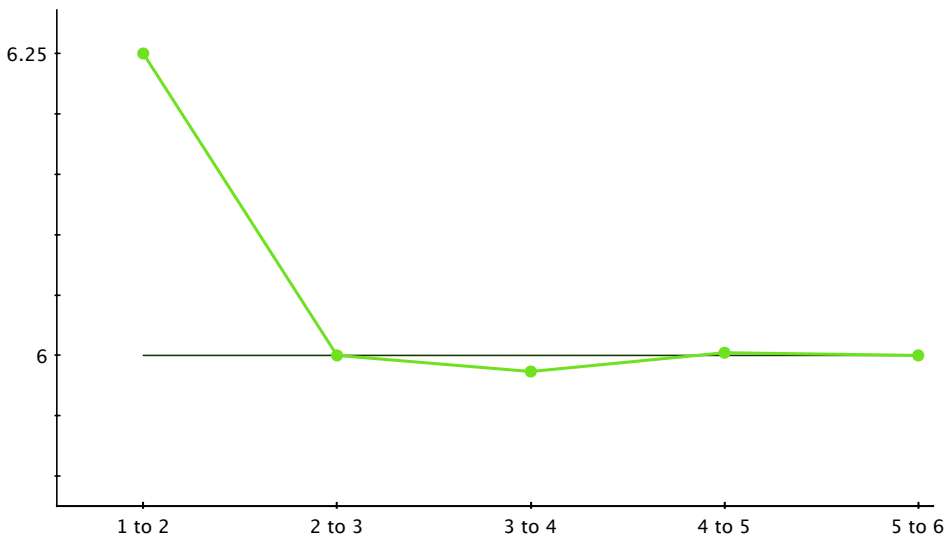
To show that mathematics can lead to paradoxes even when no infinities are involved, he reveals a curious fact that he had discovered himself. Gamblers are interested

Ergodos

in calculating the so called *critical number* of a game. It is the number of times a game needs to be repeated in order to shift the odds from the gambling house to the player. For example, if you are offered to bet on a specified side of an ordinary die, you should only accept the bet if you are guaranteed to throw the die at least four times. Then you have more than a fifty-fifty chance to win the bet. The critical number for this game is thus four.

The strange thing Gombaud had discovered was that when playing this game with two dice the critical number is not what you would expect. As the number of possible outcomes increases by a factor six when playing with two dice instead of one, the critical number ought to be increased by a factor six as well. Six times four is twenty-four, but strangely enough the critical number when using two dice is not twenty-four but twenty-five. How could this be? According to Monsieur Gombaud this fact was nothing less than a scandal, as ordinary arithmetic apparently contradicts itself. One of the other gentlemen on the journey, Blaise Pascal, got upset by Monsieur Gombaud's view of mathematics as something beautiful but poorly connected to reality, and sometimes even contradicting itself. He decided to prove Monsieur Gombaud wrong as soon as he was back at home.

Monsieur Gombaud's problem is interesting. Ideally, the critical number really should increase by a factor six for every new die we add, for the reasons Gombaud devised. However, due to the discrete nature of a die the rule is not correct for the first few dice we add. As we add more and more dice the factor does indeed approach six. See the graph below.



For this particular game it is easy to calculate the critical number exactly for any number of dice. In general, however, this is not the case. Usually it quickly becomes an impossible task due to the vexing number of combinations of outcomes that need to be calculated and ordered. The ideal property of Monsieur Gombaud proves to be a handy tool for calculating the approximate critical number for games that need to be repeated many times to break even.

Unfortunately, Pascal never solved this problem, as he did not view it as interesting. In fact, he did not even understand the question. Instead he focused on the other gambling problem Monsieur Gombaud brought up during the trip. It was an old puzzle already then but new to Pascal, known as the Problem of points.

Two persons agree to put some money at stake and the goal of the game is to collect a specified number of points, say ten, alternatingly throwing a die. If they decide to quit playing for some reason after they have started to collect points, how should the stake be divided in a fair way? If they have collected an equal amount of points the stake is simply divided in half, but what to do if one of them is in the lead?

This problem had puzzled mathematicians and philosophers for over a century with no consensus on how to solve it. Pascal started to discuss this problem with his father's friend Pierre de Fermat via a series of letters. Initially Pascal was not sure at all that his solution to the problem was correct. However, when Pascal learnt that Fermat independently had arrived at the same division, albeit using other mathematical arguments, it made a huge impact on him. He quickly became convinced not only that their new principle was correct, but also that it could be applied to any type of decision problem. For instance, he devised a novel argument based on this principle for why one ought to believe in a god, today known as Pascal's Wager.

In modern terminology, their solution to the Problem of points amounts to the idea that each player should get a share of the stake that is proportional to their probability to win the game, had the game not ended. This share became known as the expectation or the expected value. Ever since its inception, this concept has been hugely influential in a number of different human inquiries.

Fair Values

Why did Pascal and Fermat view the expected value as the fair division of the stake? Fermat apparently did not see the need for an independent justification at all. He seemed to think that it is mathematically obvious that his solution is the correct one. Pascal, however, tried to justify his solution by using a mathematical reasoning where each

player's fair share of the stake in the end can be reduced to a fair coin flip. To flip a fair coin to win your fair amount is seen as the quintessential fair game, and anything that can be reduced to this game ought to be fair as well. But to be *reduced* to a fair coin flip and *be* a fair coin flip are two different things. This is Pascal's mistake, which eventually led to the St. Petersburg problem and the messy philosophical situation we have today. However, no one at the time spotted this flaw in his argument. On the contrary, mathematicians all over Europe instead began to be interested in this new branch of mathematics, founded on the concept of expected value.

The same year the St. Petersburg problem was discovered the first version of the Law of Large Numbers was proved, giving the concept of expected value a much-needed theoretical support. It states that the expected value can be viewed as the average value for a game that is played an infinite number of times. Hence, if we play a game infinitely many times, we are mathematically justified to use the expected value as the fair prize. But if we do not happen to play a game infinitely many times, how can we justify to use the expected value?

After a large number of iterations of a game, the average value begins to fluctuate around the theoretical mean value, i.e., the expected value. Half of the time the average is above the mean and half of the time it is below the mean. This implies that if we play long enough, the probability is one half that we will end up as net winners and the other half that we will end up as net losers—if we pay the expected value each time we play the game. Just by iterating, any game will, in this sense, end up being equivalent to tossing a fair coin, which is the quintessential fair game. So, even if for a given game the expected value is not in any sense fair for a single or a few rounds, when repeated sufficiently many times, the series of rounds viewed as a whole will always be a quintessential fair game.

We see by this that even if the expected value from the outset is not a fair prize in any reasonable sense—just by repeating the game over and over we will arrive at a situation that models the quintessential fair game, i.e., tossing a fair coin once. That is, if the expected value is finite. If the expected value is infinite, as in the case of the St. Petersburg game, we will still come closer and closer to the “fair” value, which is infinity, but of course only from finite values. That is, only from ‘below.’ We will never reach a state where the accumulated gain will fluctuate around infinity, half of the time above infinity and half of the time below infinity. This is why the expected value is so strongly felt as being an unfair prize for the St. Petersburg game. It is not the infinite prize *per se* that is the problem, but the fact that the game cannot model the quintessential fair game no matter how many times we play the game. We can now define what a fair game is in the classical sense.

Definition A game is *fair in the classical sense* if and only if the probability is equally big for a net gain as for a net loss for all large number of rounds of the game.

Every game with a finite expected value can be turned into a fair game in the classical sense by assigning the expected value as the prize for the game. Note that no significance at all is put on how *much* we win or lose in the long run, only that the probability for a net profit is equal to the probability of a net loss. A net profit of one dollar is considered a 'win' as much as a net profit of millions of dollars. If we have a fair game in the classical sense we can guarantee that we will win or lose with equal probability in the long run, but we cannot, in general, say anything about how large the net gain or net loss will be. A closer analysis of the game at hand is needed for that. Hence, for the definition of fairness in the classical sense the size of the possible gain or loss is irrelevant.

However, we always need to perform a deeper analysis of the game at hand to know how *many* times we need to iterate the game in order to have reached the state when the game is fair in the classical sense. For some games the expected value is fair in the classical sense from start while for others we need to play an astronomical number of rounds. This is the meaning of the 'large numbers' in the Law of Large Numbers. But if this fifty-fifty chance for loss or gain in the end is the basic intuition behind the concept of fairness, why not use this upfront as the definition of a fair game?

Definition A game is *fair* if and only if the probability is equally big for a net gain as it is for a net loss.

From these definitions, we see that a game is fair in the classical sense only if a long sequence of games taken as a whole eventually becomes fair. But if a game is fair, it is automatically also fair in the classical sense. In the language of mathematical logic, 'fair' is a conservative extension of the old concept 'fair in the classical sense.' Hence, we have nothing to lose in adopting this new more general concept of fairness.

In mathematical language, an equal probability for gain or loss is called the median. The expected value is not a median but a probability weighted mean. What we noted above is that after a sufficiently large number of rounds played, the mean behaves exactly like a median—it actually becomes a median. This means that the median will approach the mean more and more the more the game at hand is repeated. It is in fact this median-like property of the expected value in the long run that is the only valid justification for calling the expected value a fair prize.

Solving the St. Petersburg Problem

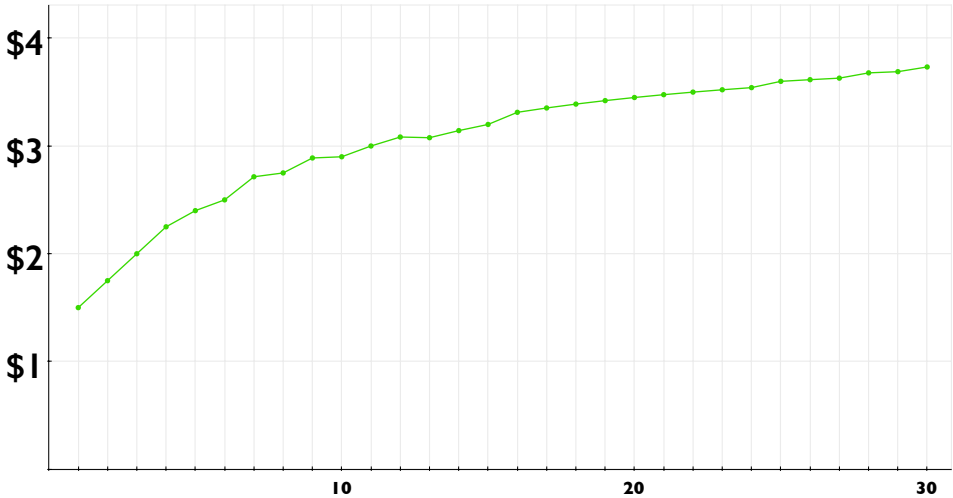
If the approach to the median is the only reasonable reason to stick to the expected value—why not use the median upfront? The median is a fair prize by definition. For any game with a uniform distribution, the median and the mean will coincide. For a game with a non-uniform distribution, the mean and the median will not coincide in general. For example, the St. Petersburg game has an infinite mean (expected value) while the median (fair prize) is only 1.5 dollars.

If we play the St. Petersburg game twice, the worst-case scenario is that we win only one dollar each time, that is, two dollars in total. This will happen with probability $1/4$, because first we have to win one dollar with probability $1/2$ and then another with probability $1/2$, and $1/2 \times 1/2$ is $1/4$. If we have a little more luck we win one dollar the first round and two dollars the second round, in total three dollars, which has probability $1/8$, as $1/2 \times 1/4$ is $1/8$. Or, we win two dollars the first round and one dollar the next round, which also totals three dollars with probability $1/8$. Adding the probabilities for all these three worst case scenarios we get $1/4 + 1/8 + 1/8 = 1/2$. So, with probability $1/2$ we will win 3 dollars or less. Hence, with probability $1/2$ we will win 4 dollars or more. The median for playing the St. Petersburg game twice is thus 3.5 dollars. The fair prize per game is therefore in this case 1.75 dollars, which is slightly more than the fair prize for playing the game only once.

That the fair value increases is what we can expect as we know that the median must approach the mean, i.e., the expected value, the more rounds we play of the game. In this case we know that the expected value is infinite. The fair prize must therefore increase without bound for larger and larger sequences of repeated rounds of the game. In practice, this means that the more we are guaranteed to play the St. Petersburg game the more should we be willing to pay for the privilege to play the game. That a fair prize must, in general, vary depending on how many times we plan to play a specific game was realized already by William Feller, as we have seen.

In the graph on the next page we see how the fair prize per game increases as we are guaranteed to play longer and longer sequences of St. Petersburg games. As already mentioned, we know that this curve must increase without bound. But can we find an expression that approximates this curve? This is in fact possible using the same idea as Monsieur Gombaud used.

The more we play the game the more certain we are that we will win the most common prizes in proportion to how likely they are. If we play four times we can be somewhat sure that we will win the one dollar prize half of the time, that is in two of the four cases. This will give us two dollars. Ideally, we would expect to win two dollars in one



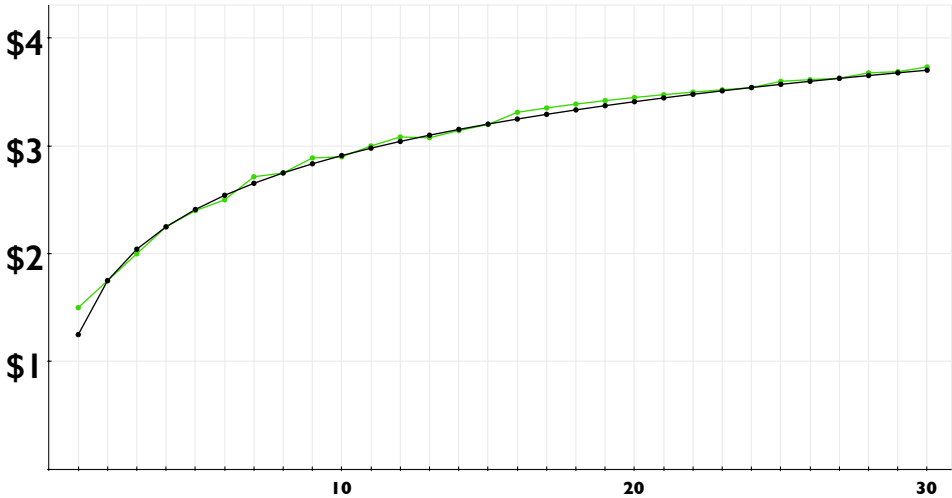
of the two remaining cases, four dollars in half a case, eight dollars in one fourth of a case and so on. However, this is clearly impossible, but the previous sentence describes exactly twice the St. Petersburg game played twice. The exact fair value for playing the standard game twice is 3.5 dollars as we already know. Therefore, the remaining two of the four cases contributes exactly 2×3.5 dollars. In total we have an approximate fair value of 2 dollars + 2×3.5 dollars for playing four times, or $1/2 + 1.75$ dollars per game.

If we play eight rounds, we will ideally win one dollar in four cases, two dollars in two cases and the last two cases is exactly like playing four times the St. Petersburg game twice. The approximate fair value is therefore 4 dollars + 2×2 dollars + 4×3.5 dollars, or $1 + 1.75$ dollars per game.

If we double once more and play sixteen times, we will arrive at an approximate fair prize of $3/2 + 1.75$ dollars per game. In general, the approximate fair prize per game when we play t times is $1/2 \log_2(t/2) + 1.75$ dollars, where the logarithm is to base two. This can be rewritten as

$$\frac{1}{2} \log_2 t + 1.25$$

For large values of t , this expression gives an almost exact approximation of the true curve of fair prizes. In the graph on the next page this curve is shown in black. Thus, for large values of t we can use the convenient formula above instead of calculating the exact fair value, which quickly becomes very complicated because of the vexing number of



cases to consider. Note how the ‘ideal’ black curve here plays the same role as the ‘ideal’ increasing factor of six in Monsieur Gombaud’s gambling problem about the critical number.

If we solve the expression above for t , we get the relation

$$time = \frac{4^{money}}{\sqrt{32}}$$

where we have replaced t with *time* to make the formula easier to remember. Whenever you are offered the opportunity to play the St. Petersburg game, try to remember this formula. Depending on the fee you have to pay to play the game you should not play unless you are guaranteed to, and have time to, play at least the number of times given by this formula. For example, if you are offered to play the game for a five dollars fee, you should not play unless you are guaranteed to play at least 182 times. If the fee is 20 dollars you will probably not have time to play the game even if you are given the opportunity to play the required amount of times.

This kind of advice sounds familiar. What is denoted ‘time’ in the formula above is exactly the same concept as the ‘critical number’ in Monsieur Gombaud’s own gambling problem. We can thus conclude that this idea is far from new. In fact, it comes natural to most people. In clinical studies where people have been asked what they are willing to put at stake for playing different games, the concept which best fits the empirical data is the fair prize, that is, the median (Hayden and Platt 2009).

Solving the Two-Envelope Problem

Unfortunately, the solution of the St. Petersburg problem does not solve the Two-Envelope problem. If we replace all expected values by fair values, the latter problem remains, as is seen in the Jailhouse Clock scenario. Incidentally, this proves that these two problems are totally unrelated. The opposite stance, that they are closely related or even just variants of the same problem, is quite common among the commentators to the Two-Envelope problem. According to this view, the true solution of the St. Petersburg problem would automatically resolve the Two-Envelope problem. Now we see that this is not the case.

The Two-Envelope problem is superficially similar to the paradoxes invented by Joseph Bertrand in the nineteenth century that helped to kill the classical interpretation of probability. However, the Two-Envelope problem goes deeper as it also shows that the Bayesian interpretations are wrong. According to Bayesian philosophy 'uncertainty' or 'lack of knowledge' can always be modeled by a probability distribution in a way that does not lead to contradictions. But we can construct an explicit probability distribution describing a certain state of 'uncertainty' for the Two-Envelope problem that does indeed lead to a contradiction (Broome 1995). This shows that the Bayesian philosophy is false.

The underlying world view motivating the Bayesian philosophy is determinism. If the world is deterministic only our 'lack of knowledge' can be the reason for why we are uncertain about what will happen. It is always, however, possible to learn more about the situation at hand and reduce our uncertainty. For example, if we flip a coin and we have no clue which side will come up, we say that chances are fifty-fifty for either side. But if we know more about the actual coin or learn the physical details on how it is flipped, it is always possible to make a better guess. If we have total knowledge of the physical situation at hand we can predict with certainty which side will come up. Hence, for determinists it is natural to equate probability with lack of knowledge. Total lack of knowledge leads to a uniform probability distribution among the possible outcomes. Partial knowledge leads to some other probability distribution, which completely describes the state of knowledge. If we have total knowledge everything is certain and we have no need for probabilities. Or equivalently, all probabilities are either zero or one. We know for certain if something will happen or not.

To use probabilities as an irreducible entity in a fundamental physical theory is thus unthinkable for a determinist. If the theory really is fundamental it cannot rely upon a concept that is synonymous to 'lack of knowledge.' That is just another way to say that the fundamental theory really is not fundamental at all. There must exist some even more fundamental theory that explains the apparent randomness. This was exactly the

view Albert Einstein held regarding the new physical theory describing the very small, Quantum Mechanics. It is held to be a fundamental theory that nevertheless relies upon irreducible probabilities.

As a determinist, it was evident for Einstein that Quantum Mechanics cannot be a fundamental theory. To convince even non-determinists he developed, together with two coworkers, a philosophical argument that is now known as the EPR argument (Einstein et al. 1935). The argument uses two spatially separated particles that are connected in a special “spooky” way that Quantum Mechanics permits. Einstein viewed the connection as spooky because the particles seem to keep track of each other through space and time in an inexplicable way. For example, if a property like spin is measured in a particular direction for one of the particles the other particle always has a spin in the opposite direction. This would not be strange at all if their spin directions were predetermined and simply set in different directions from the beginning. But that is not how it is according to the theory. According to Quantum Mechanics, the outcome of the first measurement is completely random and not predetermined at all. How the other particle can know the outcome of a completely random event far away and adapt its properties accordingly is what is called “spooky action at a distance” in the EPR paper.

As soon as one of the particles has been measured, we know for certain the outcome of the corresponding measurement of the other particle. A property that can be predicted with certainty must imply that the property is real and exists objectively, even before we measure it. But the theory explicitly denies that this property could be real and existing before the measurement. According to Einstein, this clearly shows the incompleteness of Quantum Mechanics. A complete theory has something corresponding to every element of reality. Any property in the natural world that can be predicted with certainty, i.e., with probability equal to one, is an element of reality. This is how ‘an element of reality’ is defined in the EPR paper. As Quantum Mechanics cannot account for some quantum properties, which clearly are elements of reality according to this definition, the theory must be incomplete. This is the EPR argument.

The EPR argument is flawed because the reasoning is circular. If the world is deterministic we already know that Quantum Mechanics is incomplete, so the EPR argument cannot assume that. But the interpretation of probability used in the definition of “an element of reality” is Bayesian—probability as a measure of how complete our knowledge of a situation is. And, as we know, this interpretation only makes sense in a deterministic world where every particle has well determined properties all the time. So instead of giving a definite proof of the incompleteness of Quantum Mechanics, as was the intention, the EPR argument only shows that Quantum Mechanics is incomplete if

we assume that Quantum Mechanics is incomplete.

In fact, what the EPR thought experiment really shows is quite the contrary. It shows that irreducible true randomness indeed exists. There is a well-known theorem, called the no-communication theorem, which says that the coupled particles in the EPR setup cannot be used to send information faster than light. The reason we cannot utilize the “spooky action at a distance” for actual communication is because the first measurement we perform is totally random. If it were not completely random we would actually be able to send information faster than light. Indeed instantaneously. Not only is superluminal communication forbidden in the theory of relativity, but the very concept of ‘instantaneity’ is totally alien to it. So, in order to comply with Einstein’s own theory of relativity, the EPR experiment shows that the world actually is indeterministic and not deterministic, which Einstein, Podolsky, and Rosen implicitly had assumed.

Knowing that probability must have to do with randomness and that our physical world is in fact indeterministic, does this by itself solve the Two-Envelope problem? It is an interesting fact that it does not. In the Jailhouse Clock scenario, for example, we can specify explicitly how to randomly select the pair of execution orders to be presented to you. The guards can simply use a twelve-sided die as the random generator and select one of the twelve possible pairs depending on what the die shows. Then you flip a coin to decide which of the two execution orders to choose. The Jailhouse Clock problem, however, prevails. This is why the Two-Envelope problem is deeper than the paradoxes presented by Joseph Bertrand. They all disappear when a random procedure like this is specified explicitly.

Which additional idea is needed to solve the problem? We know that the world is indeterministic because there are genuinely random events. But what does “random” really mean? In a deterministic world when repeating an experiment the outcome must always be the same. If it is not the same, we have not repeated exactly the same experiment. This is not the case in an indeterministic world. Here we can repeat exactly the same experiment and the outcome can still be different from one performance to the next. This is what we mean by fundamentally random events. So, every experiment in an indeterministic world leads to a set of random events. But is the opposite implication true? Does a set of random events imply an experiment?

Let us investigate this question in the context of the Jailhouse Clock scenario. You look at the letter with the execution order you have picked that says that you will be executed at 4 o’clock some day. Then, suddenly, you are given the opportunity by the prison guard to choose the other letter instead. This event is a random event where we do not know what the experiment is. Either the time stamp “4 o’clock” is essential to

why you got the opportunity to switch letter by the prison guard, or it is not. As this situation will only happen once in your life you have no way of knowing what would have happened if some other time stamp would have shown up in the letter you first picked.

There are two options. Either you are given the opportunity to switch to the other letter whatever timestamp you got first, or you are not. In other words, either the timestamp "4 o'clock" is part of the description of the experiment, or it is not. If it is part of the description of the experiment you are only given the opportunity to switch in case you found the execution order with timestamp "4 o'clock." In this case it is advantageous to switch as the fair value of your time left alive is increased using the other execution order. If "4 o'clock" is not part of the description of the experiment it must instead be one of the twelve possible outcomes of the experiment and you are given the opportunity to switch whatever time is stated in the first letter. If this is the case, there is nothing to gain by switching to the other execution order.

By this we see that a set of random events by itself does not imply a unique experiment. Sometimes a set of events is compatible with more than one experiment. Moreover, this also shows that attaching probabilities to events when the experiment is not defined leads to contradictions. 'Experiment' is thus a more fundamental concept than randomness.

Definition An *experiment* is a complete set of instructions on how to properly repeat something.

Note that this definition does not make any distinction between real experiments and thought experiments. It does not matter if the experiment is actually performed or not. It can occur many times, only once, or never. The set of instructions is always the same. If no experiment is defined, it is impossible to talk about probabilities. It leads to contradictions as we have seen. The experiment is thus the important concept here and not the random event itself. Random events cannot be attributed probabilities without a reference to an experiment. This is the key observation to understand the concept of probability.

Definition A *probability* is a measure of the relative occurrence of an outcome of an experiment, would the experiment be repeated indefinitely.

Note how this concept does not make any distinction between theoretical probabilities and actual measurable probabilities. This is because it relies upon the

concept of experiment above which does not distinguish between actual experiments and thought experiments.

This concept solves all probabilistic versions of the Two-Envelope problem. Obviously, Smullyan's non-probabilistic version of the Two-Envelope problem cannot be solved by a new concept of probability, as no probabilities enter that problem. Smullyan's paradox instead shows something very interesting about the logical structure of the world.

We cannot use our probability concept to solve Smullyan's problem but we can use the concept which is the logical basis for it, our concept of experiment. The two conclusions derived in the problem, that the amount we might gain is greater than what we might lose, and that the amount we might gain is equal to what we might lose, respectively, are only possible because different experiments are used to derive the conclusions. In the first case, the amount A is part of the experiment while in the second case neither Y nor 2Y are part of the experiment. In the first case, the outcomes of the experiment are 2A and A/2 while in the second case the outcomes are Y and 2Y. This is the reason we get different conclusions. Our concept of experiment thus solves Smullyan's non-probabilistic version of the Two-Envelope problem.

Consequences for the Measurement Problem

We have shown that in order to avoid contradictions we always need to explicitly specify what our experiment is and clearly distinguish that from the results we get from performing the experiment. Not only when probabilities and random events are involved, but always.

This has interesting consequences for understanding the so-called Measurement problem in Quantum Mechanics. The Measurement problem is how to explain why measurements play such a peculiar role in Quantum Mechanics. As long as we do not measure a quantum system, the system evolves in a deterministic fashion. But as soon as we do a measurement the deterministic progression collapses and we get a random result that is in accordance with what the theory predicts. The problem is that this "collapse" of the deterministic evolution is not explained or motivated in the theory at all. Moreover, no one knows for sure what a "measurement" really is. No definition whatsoever is given by the theory. It is a mystery how this concept, which is in no way a part of the theory, can nevertheless be totally crucial for the application and therefore success of the theory. To explain this is called the Measurement problem.

If we try to understand Quantum Mechanics using a deterministic ontology, we will fail, as we have seen. In a deterministic world every object has exact positions

and properties all the time. But the world is indeterministic as we have shown, so we need to adopt an indeterministic ontology instead. In an indeterministic world, there are genuinely random events that cannot be reduced to more basic underlying random events. But the only way a genuinely random event can exist is because it is the result of an experiment. If we do not have an experiment, we cannot have a random event. This is true on a basic logical level of reality for which the physical reality has to comply. Nature has solved this problem by letting quantum systems evolve in a deterministic but *unreal* manner when no experiment is defined. Or rather, between the start of an experiment and the end of the experiment. As soon as the experiment ends, which is usually called a measurement, a random event will occur that is compatible with the experiment.

If an experiment is changed, the possible random outcomes are changed as well. To investigate the state of a particle in the middle of a genuinely random experiment is impossible, simply for the reason that such a measurement changes the original experiment. An experiment cannot have two different starting points at the same time. This follows from the definition of an experiment. As soon as the particle is interrupted, it marks the end of an experiment and a new experiment begins where the previous starting point is completely forgotten.

As the change of an experiment happens at a logical level of reality, the change has to be instant. This explains the instant change in the EPR setup. Nature's detection of which experiment is at hand is always instant. If it was not instant, it would be possible to arrange a situation where two experiments were defined at the same time for the same set of particles, which would easily lead to a logical contradiction.

The new probability concept, which we can call the experimental interpretation of probability, also explains why a "certain" prediction of an outcome, as in the EPR argument, still does not mean that the particle is in that state prior to the measurement. A measurement is a physical interaction which either has occurred or has not occurred. If it has not occurred, the particle is still in its unreal but deterministic state. As soon as the measurement happens the particle's property that is measured becomes real. The fact that we can with certainty predict the outcome of an experiment does not change this fact.

Conclusions

At first, it might seem like a good thing that we allow for many different interpretations of the probability concept. But if that is the case, it must logically be a good thing to have many different interpretations of any scientific concept like force, distance, time, electric

current and so on. It is easy to see that if we had that, science would hardly be possible. There is no legitimate reason to claim that probability is special, in any sense, from any other scientific concept. Either it is good for all concepts to have many contradicting interpretations or it is good for none. It is clearly the latter case.

Allowing for only one interpretation of probability will have a major impact on both natural science and the human sciences. There will be no reason to use totally different types of mathematics in economics and physics, for example, as is currently the case. As this paper shows, both can instead start to use the same concept of probability. If adopted, this will have a great unifying effect across different disciplines.

References

- Broome, John. 1995. "The Two-Envelope Paradox." *Analysis* 55 (1): 6–11.
- Dutka, Jacques. 1988. "On the St. Petersburg paradox." *Archive for History of Exact Sciences* 39 (1): 13–39.
- Einstein, Albert, Boris Podolsky, and Nathan Rosen. 1935. "Can quantum-mechanical description of physical reality be considered complete?" *Physical Review* 47 (10): 777–780.
- Feller, William. 1945. "Note on the Law of Large Numbers and "Fair" Games." *The Annals of Mathematical Statistics* 16 (3): 301–304.
- . 1950. *An Introduction to Probability Theory and Its Applications, Volume I*. New York: Wiley.
- Hayden, Benjamin and Michael Platt. 2009. "The mean, the median, and the St. Petersburg Paradox." *Judgment and Decision Making* 4 (4): 256–272.
- Jorland, Gérard. 1987. "The Saint Petersburg Paradox 1713–1937." *The Probabilistic Revolution, Volume 1: Ideas in History*, ed. by L Kruger, LJ Daston, and M Heidelberger, 157–190. Cambridge: MIT Press.
- Samuelson, Paul. 1977. "St. Petersburg Paradoxes: Defanged, Dissected, and Historically Described." *Journal of Economic Literature* 15 (1): 24–55.
- Smullyan, Raymond. 1992. *Satan, Cantor, and infinity, and other mind-boggling puzzles*. New York: Alfred A Knopf.
- Todhunter, Isaac. 1865. *A history of the mathematical theory of probability from the time of Pascal to that of Lagrange*. Cambridge and London: Macmillan and Co.

Journal of Cognition and Neuroethics

The View from Vector Space: An Account of Conceptual Geography

Joshua Stein

New York University

Biography

Joshua Stein is pursuing an interdisciplinary Master of Arts at New York University, developing a thesis on the epistemological and social role of scope, simplicity, and precision in the formation scientific theories, under the advisement of Michael Strevens. He is also developing projects on the psychology of concepts and its normative implications for philosophical methodology and a paper on the history of disputes between literary theory and analytic philosophy of language. His past work ranges in scope from multiple realizability and its role in comparative psychology to the differences between pedagogical and disciplinary styles in martial arts schools and secondary and post-secondary education.

Acknowledgements

I am grateful to, among others, Arnon Cohen, Brent Kiou, Benedicte Veillet, and an anonymous commenter for clarifying the areas of the expository section.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Stein, Joshua. 2014. "The View from Vector Space: An Account of Conceptual Geography." *Journal of Cognition and Neuroethics* 2 (1): 73–93.

The View from Vector Space: An Account of Conceptual Geography

Joshua Stein

Abstract

This paper offers an introduction to the alternative views of mental representation put forward by the philosophers Paul Churchland and Teed Rockwell. These views differ from conventional accounts of representation in their description of mental representations as instantiated in a vector space, a particular sort of state space, where the neurological structures of the brain act as a map of, and serve to instantiate, the relevant regions of that state space. It draws a contrast between approaches to understanding mental representation in a sort of compositional semantics.

It then argues that the discrete approaches advocated by Churchland and Rockwell help to address compellingly some very difficult issues in the philosophy of mind, and help to bring both a precise account of the instantiation of certain mental states based on activation patterns in the brain and draw attention to the embodiment of these mental states. In doing so, the vector space account can be pressed into service of many important discussions in contemporary psychology. A late section of the paper prefaces the application of a vector space account of social cognition to disputes about processes of mental state ascription in comparative psychology.

It shows that reasoning is a process of enriching a representation, occurring as a result of the subsequent activations of particular pathways in the vector map. As activation patterns move through various regions of the brain, representations are modified, enriched in ways that we identify as reasoning about concepts. These rich representations of the world, with the complex social and conceptual detail, are the result of the processes Churchland and Rockwell describe.

Keywords

Representation, mental content, vector space, reasoning, social cognition, embodiment, extended mind, comparative psychology

The paper focuses on the account of mental representation put forward by Teed Rockwell and Paul Churchland;¹ it argues that, despite significant differences between the two accounts, there are important shared features that tie the two together. In tying the two positions together, it develops an argument for conceiving of mental space as vector space, and of neurological structures as a map of that space. Further, it illustrates the substantive philosophical and scientific success held by both theories, e.g., the successful

1. While large portions of this discussion are drawn from earlier work of both Rockwell and Churchland, the views are laid out fairly fully in their recent work. Churchland has discussed this approach for several decades, but most recently in *Plato's Camera* (2012). Rockwell lays out his account in *Neither Brain nor Ghost* (2007).

handling of learning histories and comparative psychology. I maintain that the differences between Churchland and Rockwell are neither entirely empirical nor philosophical; rather, the theories are successful in part because empirical differences bear on philosophical differences, and vice versa.

1. Prefacing Churchland & Rockwell

Before moving into a positive account of the views presented by Churchland and Rockwell, it is helpful to characterize the position they emphatically reject: the more common account of mental content given in the literature on philosophy of mind from the 1960s onward.² Characterizing this family of views, itself, could be an exhausting project, but I will just give a brief preface to the relevant facets of the literature. These are the ‘semantic content’ views, which maintain that the content of a representation is the truth-conditions for that representation, like content of a sentence in compositional semantics. The representation of my reclining on a brown coach is specified by conditions where “Joshua is reclining on a brown coach.” is true and false. These are the conditions whether the representation is attached to a belief, desire, imagination, etc. The content of a representation are specified the same way as the contents of a linguistic token. Some members of this family take the account literally, suggesting mental content is given in an actual language of thought with definite syntactic structure, like Fodor’s *mentalese* (Fodor 1975; 1990). This family of views has been defended extensively, with interpretations ranging in empirical and conceptual commitment; many suppose something weaker than a language of thought, but maintain a semantic view of mental representation (Perry and Barwise 1983; Searle 1984; 1990; 1992; Stalnaker 1984). For a brief overview of the points of dispute of the language of thought hypothesis, see Ayede (2010, §5) bearing in mind that the views of Rockwell and Churchland are an idiosyncratic variety of what Ayede calls “functionalist materialism.” The semantic content family concerns what mental representations are (Siegel 2012, 7–8; 21–23), rather how they come to have the character they do (Prinz 2000), but positions on the latter bear on the former. Indeed, one purpose of the semantic contents views is to give an account of the formation of mental content, physically and computationally (Rey 1992, 289; Haugeland 1985). This is also the goal of the vector space view, as we will see.

2. I am thinking here of the work following Chomsky, including work following Fodor on language of thought, laid out at length in Fodor (1975; 1981) and in brief in Ayede (2010). Fodor’s view is notably folk-psychological, in the sense derided by Churchland and Rockwell, but there are various strains of the view that are not, e.g., those developed by Carruthers (2006) and Dennett (1986; 1998).

There are a few reasons that Churchland and Rockwell reject the semantic content family of views, but one is particularly relevant to our discussion. The semantic content views attempt an assessment of mental content *per se*, independent of the physical structure that conditions its instantiation. Rockwell writes,

... advanced neuroscience will not just give us more information about what the brain does and how it does it. It could also end up eliminating the whole concept of brain, just as easily as it could eliminate any other concept originally derived from folk psychology.

Fodor, [the exemplar of the semantic content family] however, is quite explicit in this commitment when he claims that psychology must accept what he calls “methodological solipsism” (Fodor 1981). What he means by this is that mental states must be studied as an independent system that takes entirely within a brain, which can be understood without any reference to the outside world...

I call this myth... the myth of “the machine in the machine.” It is the basis for Fodor’s “language of thought” and any other theory of mind that holds that all we need to understand the mind is to open Skinner’s black box. (Rockwell 2007, 10)

Churchland has a similar view (in an omitted portion of the passage above, Rockwell says that Churchland “almost breaks free”) in the consideration of a form of functionalism “construed broadly as the thesis that the essence of our psychological states resides in the abstract causal roles they play in a complex economy of internal states mediating environmental inputs and behavioral outputs” (Churchland 1989, 23). Churchland’s account of vector-spaces is offered to provide a very specific, neuroscience-driven account of how the environmental inputs realize complex representations that allow for sophisticated behavioral outputs.

1.1 Churchland’s account of vector-space & activation pattern

Churchland’s alternative to the semantic content views is to account for mental states as neurological activation patterns (2002, 27–32; 1996, 21–36). The potential activation patterns in a given region of the brain map a vector space, and activation specifies the content of a tokened representation. The structure constitutes a state-space in which a given state of affairs can be represented. The neurological structure constitutes the

standing possibility of possible representations i.e., the state-space underlying any and all possible representations, and when the perceptual system is stimulated in a certain way, those neurological structures activate in a pattern specifying the content of the representation. Churchland arrives at something very much like an identity claim about the relation between mind and brain, where the vector space and the vector-map collapse together. This leads to the controversial claim that mental states simply *are* activation patterns, rather than discrete entities.³

Any sense in which there is some predicate content of the sort postulated by a classical content view is, then, capable of being given in terms of a vector coding (Churchland 1996, 29-30; 1989, 154-196), and content given in the vector space account is capable of being ascribed truth-conditions. What Churchland is really after is a story about how the neural structures realize rich representations. The general account is straightforward. He starts his own account,

...at the bottom of several ladders, each of whose bottom rung is a population of *sensory* neurons—such as the rods and cones in the retina of the eye, the ‘hair cells’ in the cochlea of the ear, or the mechano-receptors within the skin. Each of these initial neuronal rungs projects a wealth of filamentary axonal fibers ‘upward’ to a proprietary second rung—to a receiving population of postsensory neurons... each *rung* of these ladders constitutes a unique cognitive canvas or representational space, a canvas or space with its own structured family of categories, its own set of similarity and difference relations, and its own peculiar take on some enduring aspect of the external world. What happens, as sensory information ascends a ladder, is its progressive transformation into a succession of distinct representational formats, formats that embody the brain’s background ‘expectations’ concerning the possible ways in which the world can be. (2012, 35)

Consider one of Churchland’s favorite cases: facial perception. Churchland (*ibid*, 7–10 and 62–74) gives a fairly straightforward account of how neurological activation patterns

3. There are a number of important philosophical objections. The first is the standard line raised against an identity relation given by Chalmers (1997). I do not think that this objection turns out to be as devastating, even to someone who is sympathetic to much of Chalmers’ account, since we can say that the identity relation is not a matter of the intension, but a matter-of-factual claim about the relation in our world; this response requires additional fleshing out. The second objection comes from Bennett and Hacker (2003; 2007). I take this objection to be answered by methodological and conceptual criticisms advanced by Dennett (2007).

specify the particular details of a face and allow for facial recognition within the state-space, showing that the neural network is “trained to discriminate faces from nonfaces, male faces from female faces, and to reidentify the faces of various named individuals across diverse photographs of each” (ibid, 7). The multi-dimensional state-space realized by laddered neural networks allows for representation and, through weighting of certain collections of neurons throughout the system, memory of a complex array of faces. This offers a theory of facial perception and representation, and lends itself well to projects in mapping the relationships between neurological and psychological states.

The view can be extended even further, offering an explanation of the representation of salient social features, like facial expressions. When a subject sees someone experiencing distaste, there is the presentation of the visual stimulus of the face to the visual system, the categorization of the face as such in the activation pattern, and then the sequential activation of a series of regions and subregions associated with features of the facial recognition vector space, i.e., the progression of rungs. Like recognition of a face as a particular person, recognizing the expression on someone’s face is a higher rung on that ladder. The behaviors associated with the ascription of a particular mental state are recognized by the activation of certain regions of the vector space, and those regions are connected to other regions related to other features of the cognitive architecture, e.g., simulation of affect, or whatever consists mental state ascription; the cascades set off by the initial presentation of the visual stimulus to the visual system are sufficient to explain the progressive enrichment of the representation, e.g., the recognition of the face, the association of that face with a particular person, the ascription of mental states to that person, etc. The subject ascribes distaste based on the progression of the activation pattern representing the stimulus, explaining the defeasibility of judgments about mental states when certain features of the activation pattern are incidentally activated by a stimulus, like the phenomenon of representing faces in Arcimboldo’s paintings of non-faces.

As both the semantic and vector space accounts note, there are a number of functions tied up in facial representation pertinent to the derivation and ascription of emotional responses. These features are often subtle and not obviously part of a semantic operation, as they are not given as propositional. Particularly tricky for semantic content views are motor function and affective processes mapped to limbic activity (Gopnik and Seiver 2009; Hutto 2012), as well as visual attention to features of the face, which explain heterogeneous results for tests of facial recognition among subjects with Autism Spectrum Disorder (Hobson and Hobson 2012). Because these features require an account of embodiment, of the way that neurological features interact with parts of the peripheral nervous system and sensory organs, describing them as propositional attitudes

is problematic.⁴ The semantic content views yield a sketch of social cognition where rendering a representation is a straightforward inferential process, distancing themselves from a discussion of the relevant physiological and behavioral differences in the subject. (We will revisit the importance of this difference when we talk about reasoning later.)

The major success of the vector space view comes directly from its rejection of semantic content views; by refusing to describe mental states as abstract entities with content given as truth-conditions, the vector space view can focus on embodiment when required. It can turn to the relation between sensory organs and the ladders of neurological structure that participate in representation, offering a mechanistic account of the content. If the content is handled more easily as an abstract object, we can render a heuristic account of that content which allows for that treatment. On the other hand, we can also provide a fine-grained account, a “highly specific and very different [unit] of representation” (Churchland 2012, 5) embodied within an organism embedded in an environment.

1.2 Rockwell’s account of mind as embodied and embedded

Rockwell’s account does not drift too far from Churchland’s, either in the spirit of characterizing the mental as embodied or in the vernacular used to tell the story. I trade on the linguistic similarities to reinforce the similarity of the accounts, but the deeper conceptual similarity can be demonstrated simply by looking at their goals. Churchland’s goal is to offer a fine-grained account of mental content based on the neurological structures of the organism itself. Rockwell is not so interested in this process of specification, but rather in acknowledging the role of the environment in instantiating mental states; he gives a conservative version of the claim, “even if there are or could be times when the brain is having experiences without receiving stimuli from the outside world, at least some of our experiences would not be recreated even if all of the appropriate neural activity were activated” (Rockwell 2008, 65). We need the peripheral nervous system and sensory organs to tell the story of our mental states. We need our embodiment to explain the mental representations that we have and we need the world to help bring about those representations.

4. For discussion of social cognition in particular, see Goldman and de Vignemont (2009) and Niedenthal et al. (2005). For general discussion of the role of embodiment in perception and cognition, see Noë (2006) and Lakoff and Johnson (1999). The latter two, particularly Noë, are exemplary of what I call the “mind-as-joint-relation” camp.

I suggested above that Churchland's account requires embodiment; this is true. The first step in the story that Churchland tells is at the level of sensory perception. But the goal of Rockwell's account is to bring the embodiment of the organism into the foreground, to show that (both conceptually and empirically) we cannot tell the story simply by giving a description of the brain states, nor by accounting for abstract functions. Rockwell uses embodiment to emphasize the role of the world in facilitating mental states. Because we are organisms perceiving, the act of perception needs an object; it needs a world perceived. Another way, the sort of embodied organism that we are must be embedded in a world.

Rockwell prefers the language of vector spaces (Rockwell 2008, 55) for talking about the specification of mental content. Though his language is ambiguous about the status of representation, and the use of the term "content," the view is interested in the same phenomena (Rockwell 2007). Because of the heavy emphasis on embodiment and embedding, Rockwell takes "representation" as the psychophysical states activated by sensory inputs and governing motor functions. This locates him firmly in the "mind-as-joint-relation" camp in northern California, among George Lakoff, Alva Noë, Bert Dreyfus and others.⁵ The entanglement of representation with other cognitive functions makes this view appealing, because the view demonstrates the continuity of direct perceptual experience (the stimulation of the receptor cells) with the peripheral and central nervous system. It does not require an attempt to discretely identify cognitive functions. Instead, it just takes "representations" to indicate the role of some set of functions within the system.

The neuroconstructivist view advanced by Rockwell is similar to the view of brain plasticity advanced by Churchland (1979), because both have an interest in the serviceability of epistemology. They want a theory of representation that can account for some representations as instances of knowledge. Both use "attractor" to indicate a series of vectors activated together, changing the vector space over a period of time (Churchland 1989, 207–208). This neurological and mental change is the basis for an account of learning. Like Churchland, this view of learning allows Rockwell to tie

5. Those who are particularly historically astute will note that some of the primary contributors to this tradition are not a part of the immediate geographical region, among them Edwin Hutchins (once colleague of Churchland at UC San Diego) and Mark Turner. The geographical sketch is a rough heuristic for identifying communities, recognizing the immediate influence of communities of philosophers on each other. Eliminativist discussions have played a greater role, generally, in southern California thinking on mind (even amongst sympathizers with embodied psychology, e.g., Antonio Damasio) and embodied cognition has been definitive of the discussion of philosophy of mind in much of northern California, especially around UC Berkeley.

representation into other cognitive functions, instead of giving an account of reasoning as the application of abstract functions to tokens of propositions, like the semantic content views do. For the eliminativist, we come away learning history as neurological change, creating simple causal stories about precisely how neurological structure changes between the initial state and the learned state; for the mind-as-a-joint-relation camp, the story admits of more conceptual complexity, because learning involves the presence of a certain sort of stimulus, i.e., one that acts as justification, maintaining the philosophical difference between learning and the fostering of delusions. The two accounts experience tension (and the two advocates experience outright disagreement) in the desire of the eliminativist program to focus heavily on the brain-state and frustration with that approach within the mind-as-a-joint relation camp, but they can be conjoined so the fine-grained account of representation offered by Churchland fits neatly with the extension of the mental state into the body and world.

2. Points of Incompatibility between Churchland & Rockwell

The views, like the regions where they emerged, can appear to rather close to outsiders, but disparities should not be understated. They emerge out of very different approaches to philosophy, and are independently attenuated to conventional claims in the contemporary philosophy of mind. I will come to the former a bit later. My intention is to show that the general philosophical approach turns out to be largely irrelevant to the stories they want to tell about mental representation and, thus, to my re-appropriation of those stories. The burden is on me to show that programmatic baggage can be divorced from the account. With this in mind, a number of the important philosophical platitudes are immediately available for reflection and evaluation, and their role in these accounts helps establish the contrast that needs to be addressed.

2.1 neurocentrism & the empirical assessment of the peripheral nervous system

Consider the claim explicitly disputed by the “mind-as-joint-relation” philosophers like Rockwell: the claim that mind is a function of/caused by/identifiable with the brain. The mind is in, or emerges from, the space between our ears. I will call this view ‘neurocentrism.’ Rockwell says that neurocentrism is false; it is an open empirical and conceptual question as to whether the mind is even subdermal. Far from being an obvious platitude used to evaluate whether an account of mind is plausible, this claim faces a great deal of serious argumentation, especially on the basis that the peripheral nervous system and the sensory organs play an important role. Consider a counterexample to

neurocentrism in sensorimotor coordination and embodiment. It seems that sensorimotor coordination occurs largely in the peripheral nervous system (Rockwell 2010, 734–744), in parts of the system that respond directly to stimuli rather than requiring input from the neurological system; e.g., a sharp or hot sensory input occurs in the hand, and the peripheral nervous system proper responds by contracting the arm to pull the hand away; the sensory input will still reach and be represented in the brain, but the actual task is being carried out in the peripheral nervous system and so a substantial portion of the mental state occurs in the arm, not in the brain.

While Churchland has argued that much of our sensory motor coordination is based in the central nervous system (Churchland 2002, 26–27; 1996, 91–96) both he and Rockwell take this to be largely an empirical issue. Which systems are responsive to stimuli that cause disorientation? Which are activated in the performance of certain tasks e.g., regaining balance after a slip? These questions are not a matter of introspection on the phenomenal character, but rather about the empirically assessment of the central nervous system in the performance of these tasks. “Mind-as-joint-relation” advocates, including Rockwell, take this as an empirical point in favor of the notion that neurocentrism, as advocated by Churchland, is false.⁶

Why should we care about the acceptance or rejection of this platitude? Partly because it is not an accident that one account challenges a platitude while the other accepts it. Neurocentrism is central to Churchland’s views because of his account of developmental psychology and plasticity of mind, not simply as an incidental feature derived through the recitation of the platitude by his peers. Roughly, it is important to Churchland that the child’s perception of the world be minimally mediated in order to lend itself to the reliable representation of the world that he is after in his epistemology of science. If embodiment plays a major role in facilitating the development of the brain, then this calls into question whether the development of perception and representation tracks the truth. Churchland is not enormously dogmatic in his espousal of the view, though, so long as complex representation develops as a result of the causal power of the world in a way inclined to track the truth of the situation, something that an advocate of the mind-as-joint-relation view can grant without problem.

6. There are a number of arguments for extended mind. In the case of Rockwell (2010; 2007) and Noë (2010), this claim is taken to be empirically informed. By contrast, there are conceptual arguments for the position, most notably Clark and Chalmers (1998).

2.2 Intersubstitutability of mental & neurological states

Another relevant conceptual issue is the disagreement about the non-intersubstitutability of mental and neurological states, where it is Rockwell who assumes the orthodox position. The mainstream view is simply that, whatever is constitutive of (or identifiable with) a mental state cannot be the neurological state. They are not the same. Therefore, in our description of a state, it is not acceptable, not sensible, to substitute a neurological state with a mental state. Churchland's denial of this claim is well known, as are the conceptual arguments supporting the mainstream view.⁷ Rockwell denies intersubstitutability on the basis that mental states are a joint relation, and so brain states cannot be exhaustive. A mental state includes the neurological state, as well as the state of the peripheral nervous system, salient features of the world, etc.

This is a difference in the conventions governing the terms for mental states, in large part resultant from what they are choosing to focus on in their assessments. Churchland is focusing on the neurological structure, and so it is unsurprising that he advocates for using the language of mental states. Rockwell's entire account emphasizes the importance of the joint relation, and so the fact that he should be so insistent on rejecting intersubstitutability is unsurprising, because he wants to continue to emphasize the importance of embodiment and embedding, and accounts allowing for intersubstitutability generally ignore those features. The fix is simple. For the purposes of one discussion or the other, we might be inclined to note the intersubstitutability or not, but when we perform the substitution, it is always heuristic rather than an exhaustive description of the state. We can give an account of the activation pattern as an account of the mental state, but recognize that the account is not exhaustive because it does not include the full account of that state, including embodied features and the role of the environment.

2.3 The meta-philosophical difference

These differences between the two accounts raise an issue; how are we to reconcile two accounts so different in certain parts of the story that they tell? The position that mental states compose a joint relation between brain, body, and world cannot be

7. See Kripke's (1980) and Bennett and Hacker's (2003; 2007) arguments against mind/brain identity theory. We might suppose that arguments against identity of mental states with qualia (Chalmers 1997; Jackson 1998) or intentionality (Searle 1980; 1983) might apply, but those cases are somewhat more contentious, because Churchland can deny that those phenomena occur as characterized in the argument, and that the actual phenomena are conceptually reducible. The general accounts are a much more explicit challenge.

clearly excised, nor can the claim that the underlying neurological state is an exhaustive explanation of a given mental state. This difference matters a great deal, especially if spatial perception (and, derivatively, representation) is a candidate for embodiment in the way sensorimotor coordination is. All phenomena in the vector space account of representation require an answer as how they are embodied before being able to tell the story. I maintain the best solution here is to use criteria shared by both philosophers, and they will tend to converge on the same answers, at least around the account of representation.

It is important to note the substantive meta-philosophical divide between the two accounts, in order to get at how we can justify (independently) a form of pragmatism from each position. We begin with the realism of Churchland and the pragmatism of Rockwell, with a pair of methodological choices far from exhaustive of our inclinations. In these cases, certain resolutions are satisfying for both camps, though there may be some cases where no such resolution can be reached. What I need is a set of shared criteria that allow for the adjudication of a substantial number of these cases. The shared set of criteria is addressed in Rockwell (1995), what he calls “pragmatic pluralism.” This occurs as a criticism of Churchland, based on Churchland’s acceptance of pragmatic pluralism, but failure to be sufficiently pluralist about folk psychology. In “pragmatic pluralism,” we entertain various theories and evaluate them based on a set of pragmatic criteria, entertaining theories that function with respect to some criteria in certain situations, but not dispense altogether with theories that fail to meet other criteria.

In these shared criteria, we find a strong prospective solution to the issue; with respect to accounts where the mind-as-a-joint-relation theory seems pragmatically viable (e.g., epistemological issues, reference of mental representation, etc.), we are welcome to employ the account. In cases where we want a fine-grained account, where particular features of the neural map are described, we find the eliminativist account more valuable because of the focus on the relation between the activation pattern and the state space. Those on either side might argue that there are simply not cases where they are incapable of telling the relevant sort of story; after all, surely an eliminativist can give an account of a particular epistemological state and someone with a mind-as-a-joint-relation position can give a fine grained account of a particular mental state. However, we are still inclined to prefer one position or other for offering such an account, to create emphasis, to avoid a problem, and the pragmatic pluralism accepted by both Rockwell and Churchland entitle us to do so. The appropriate solution is to treat the philosophical disputes as points of dispute about which we are non-committal, and qualify the account.

3. Reasoning as functions performed on representations

Perhaps the most contentious part of this account is the notion that this has something of interest to tell us about reasoning, and features of cognition generally understood as non-representational; this is partly because it is a departure from the excursus on Churchland and Rockwell and a venture into a different part of the sandbox. I no longer limit myself to contending that the activation pattern is constitutive of a representation, and its occurrence in a particular region of the vector-map is sufficient for an entire story. Residual of Churchland's account is the process by which the representation is modified, and it is useful to talk about that process; its variance across populations of individuals is of interest to developmental and comparative psychology, as well as the general assessment of cognitive variation in cases like Autism Spectrum Disorder.

As Churchland and Rockwell note, there are regions of the vector-map corresponding to features of a given representation, and an activation pattern in that region of the map corresponds to a representation of a particular region in the vector space. This description accounts for the rich representations that interest us. In the case of someone who has a lesion and becomes agnosiac, the lesion precludes the activation of the region of the vector-map corresponding to the feature lost by the agnosiac patient, and this account seems exhaustive. However, there are cases where we recognize the cognitive impairment as the inability of the brain to perform a function. Cases of agnosia might be given as a lost ability (e.g., the inability to see faces) though the presence of this locution inclines us to talk about the loss of a cognitive function. We can tell the story of a cognitive power ("being able to see faces") or of a representational function; it makes no difference. The cognitive capacity of facial recognition given in vector spaces, Churchland notes,

... embodies an elegant solution to a chronic problem that bedevils the classical or lingua-formal approach to abductive inference generally... What will count, for a given individual in a given situation, as the *best* of the explanatory takes or interpretations available, is a global function of *all* (or potentially all) of one's current background information. But short of conducting an exhaustive search of all one's current background beliefs and convictions—an effectively impossible task, at least in real time—how can such an evaluation ever be relevantly or responsibly performed? ... [citing Fodor 2000] the problem here at issue takes center stage as the potential Grim Reaper for the hopes of the classical Computational Theory of Mind...

Where, if anywhere, does the network's acquired background information reside? Answer: in the acquired weight configuration of the 327,680 synapses meeting the second rung, and in the further weight configuration of the 327,680 synapses meeting the output layer—roughly two-thirds of about a million information-bearing connections, even in a tiny model network. (Churchland 2012, 68)

In the case of facial perception, the cognitive powers associated with social cognition and facial recognition are functions being performed on representation in virtue of the background information provided by previous events altering the weights of the relevant collection of synapses. The loss of some portion of those synapses then restricts the possibilities for representation and so coincides with the loss of cognitive function. Now, it seems unlikely that either Churchland or Rockwell would ever deny that some things are cognitive functions, like the working memory and prediction, but it is clear on this account that these cognitive functions are, in fact, parts of the representational process given in terms of vector spaces. The purpose, here, is to illustrate cases where something can be described both as a region of the space and as a cognitive function, like the ascription of mental states. Representational capacities are themselves mental functions.

This gives us access not just to a useful locution, but also puts us in a position to characterize behaviors associated with the loss of cognitive functions.⁸ In giving an account of mental states, the functional account helps establish the relation between the neurological function of the region of the vector-map and the relevant set of behaviors correlated with that map. It helps us to tie together the mind and world in a way friendly to their conceptual employment in those cases, without taking away that fine-grained assessment that emerges from the vector-space account of mental representation.

3.1 Applications to comparative psychology

One promising empirical project, though I will only preface it here, is to develop a way of talking about heterogeneity in representational vector spaces, as opposed to talking about the presence or absence of a particular faculty in a given individual. Consider

8. One of the problems in characterization and ascription of mental states is the role of behavior. Part of the value of the mind-as-a-joint-relation account is that it acknowledges this difficulty as central in most areas of philosophy of mind, refusing to give in to a characterization of behaviorism or psychologism. The eliminativism advanced by Churchland becomes ultimately psychologistic primarily by virtue of its neurocentrism, i.e., acts of mental state ascription are primarily about internal states. If that neurocentrism turns out to be empirically defeated, then it seems reasonable to suppose that behavior will play more of a role.

“theory theory” in developmental and comparative psychology (Tomasello 1999a; 1999b; Trevarthen 2012; Gopnik 2009). These accounts, whether given in terms of attentional features or theory-of-mind, treat social cognition not as a mode of representation, but as a function performed on certain mental representations. In showing the way that functions operate in the vector-space account, we see that this literature is not lost, but impacted by the difference in approach. The alternative is to recognize a difference in the vector space, the space of possible representations, conditioning the differences in mental representation and cognition in these cases. We shift talk about social reasoning to a set of relations between regions representing features (e.g., vector-spaces for facial representation, spatial representation, temporal representation, etc.) treating some functions as enriching the representation in different ways, e.g., processing representations about affect. These processes can, but need not, be given a functional locution, despite being a part of the representational process, as characterized above. What is the upshot?

The upshot is similar to the pragmatic pluralism in the discussion above. There are instances where different approaches turn out to be valuable, either for giving an account of the mind as a computational entity or for looking at correlated neurological structures in the neurological map. There are instances where macrostructural comparisons that look at a set of representational regions need to be more coarse-grained than the vector-space account will allow.⁹ However, the vector-space account can serve to correct our locution even in these cases. When we talk about the differences and similarities in neurological and mental states between two individuals (or two populations), we often fall into the trap of talking about one group as possessing a particular cognitive capacity and another group lacking the capacity, e.g., “Horses do not have color vision.” “Sparrows cannot introspect.” etc. Where there is a major structural difference, these claims reflect the state of things, but in many of the interesting cases, it leads us to dramatically overstate the differences, or even misrepresent them.

One of the major points in comparative psychology is whether the “theory-of-mind” is present in chimpanzees or not, as though this has a “yes or no” answer. It is either there or it is not, supposing that a substantially weakened “theory-of-mind” would not count. This leads to extended and frustrated discussions in the literature in psychology

9. Carruthers’ (2006) revision of massive modularity is a good candidate for such an account. It provides a compelling and empirically viable approach to discussing differences in the cognitive architecture of various species on the basis of comparison of modules (redefined to be less objectionable than Fodor’s initial account) that seem commensurable with this account. Treating the populations of neurons as systems themselves allows Carruthers to keep an eye on the proverbial forest instead of getting lost in the complexity of the neurological microstructures.

and primatology about whether the evidence supports one conclusion or another; see ongoing disputes between Povinelli (et al. 1997; et al. 1992) and Leavens (2012; et al. 2005; et al. 2004) *inter alia*. On the one hand, Povinelli maintains that the theory-of-mind is not present in chimpanzees in the same way that it is present in young children; on the other hand, Leavens maintains that the theory-of-mind is present. This appears to be a straightforward empirical issue, and is treated as such by psychologists, but it is conceptually more complicated when we peer into the particulars. After all, Povinelli's view is not that there are no social cognitive capacities at work in the chimpanzee, and Leavens' view is not that the social cognition of adult chimpanzees is the same as the social cognition of humans. Rather, the issue is that there are substantive cognitive differences in the chimpanzees, and the extent of those differences can be assessed in part through an analysis of the social behavior of the chimpanzees.

In the initial discussion of Churchland, I gave a brief sketch of how social cognition regarding mental state ascription can arise from the facial-recognition system. We can see that an account of the socially rich representation can be given in the vector space account, but consider that we could give two parallel analyses; a generic analysis for humans and a generic analysis for chimpanzees. In comparing the two analyses, what we find is not a comparison in terms of straightforward presence or absence of a particular mental function like the ascription of a complex mind, but rather the finer differences in the social cognition of the two communities, in how the enriched representations tie in to the affect of the individual and how the affect manifests itself behaviorally. These differences are cognitive manifestations of neurological and physiological differences; some of these differences will result from the environment in which the individual members of the community develop and some will be the result of genetic disposition.

The concerns expressed by Leavens about the developmental differences and the role of environment in facilitating psychological difference is a point of open empirical discussion that lends itself well to the literature on embodiment and environmental embedding invoked by Rockwell. Given the claim that a particular feature of the environment, e.g., growing up in a community that uses some linguistic communication, is relevant for the development of certain capacities, we use empirical discussions in developmental psychology to assess impact on human and chimpanzee cognition; then account for the relevant cognitive similarity and dissimilarity at various stages of development. We maintain the more fine-grained analysis per Churchland, but bringing the causal power of the world and the embodiment of the individual in facilitating the development of social cognition into view per Rockwell. The vector space approach outlined above allows for a treatment of the cognitive psychology of both communities

in a fine-grained way, engaging with causal structures that interest developmental and comparative psychologists.

The purpose of the discussion is to turn us towards the different cognitive features, which allows us to maintain the empirical discussion that Povinelli and Leavens and co. intended to have, and it still about the same cognitive explananda as before, though framed differently. It deprives us of the more coarse-grained straightforward denial or acceptance, instead working from the shared assumption that some social cognition is present, that social cognition is substantively different, and that it is the structure of those differences that is of interest to comparative psychology. We deprive the discussion of little more than hot rhetoric, moving away from an all-or-nothing ascription of function in favor of a comparative analysis.

4. Closing notes

The purpose of the discussion is to illustrate a theory in the philosophy of mental representation. Prefacing the bridges into psychological discussions is important, as the adjacent discussions are a place where the vector space theory comes together with the embodiment and environmental embedding of the organism, demonstrating the synthesis. The commentary on developmental psychology is, at least for Churchland's purposes (and mine), more valuable than the attempt at conflict resolution. There are a number of important conceptual and empirical issues tied up in the general story of representation, and even those who dismiss the accounts given here can make use of how a dismissal positions them on a number of issues; what is at stake in the rejection of the view.

In the course of this discussion, I have addressed substantial empirical and conceptual disputes between Churchland and Rockwell, in order to bring together their accounts of vector space. The philosophical differences, both in their disagreement on issues in the philosophy of mind and epistemology, are reconcilable, as are the programmatic differences that inform the animus of the conflict. I have shown that the first set of differences are likely subject to resolution either through empirical work, or some simple conciliatory conceptual work, and that the larger conceptual conflict does not serve as a major block, because they have shared criteria for the evaluation of many of the claims under consideration. Following that, I addressed the implications of the shared account of representation for a few areas in the philosophy of mind, illustrating with the question of social cognition in comparative and developmental psychology.

Most importantly, I have illustrated the role of reasoning in representation on the vector space view, where reasoning is the enrichment of representation by moving up the ladder of the activation pattern. This models the explanation of the representation of straightforward states, like the presentation of color and the contours of faces, as well as the problematic inferential states that have frustrated prominent philosophical accounts, including and especially the semantic content views I drew on for contrast. The complexities of distinguishing faces and non-faces and ascribing mental states to other individuals can be realized through extrapolation of the vector space account; the account can be fleshed out further by extending the account of mental states out into the perceptual organs, accounting for the origination of the activation pattern in the joint between the body and the world.

The conclusion emerges from that final look at the prospective role for the vector-space account in understanding cognitive psychology. The moderate and naturalized stories offered by the vector space account of representation suggest this approach is promising for empirical psychological work. While, for Churchland and Rockwell, the concepts function as a basic philosophical foundation to address conceptual issues, they can be pressed into service in favor of the resolution of scientific problems, like those in comparative and developmental psychology, and developed into more sophisticated scientific devices in the course of that discussion. This approach promises a valuable asset in a complex and conceptually difficult workspace.

References

- Ayede, Murat. 2010. "The Language of Thought Hypothesis." in *The Stanford Encyclopedia of Philosophy*.
- Barwise, Jon, and John Perry. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.
- Bennett, Maxwell and Peter Hacker. 2003. *The Philosophical Foundations of Neuroscience*. New York: Wiley Blackwell.
- . with John Searle and Daniel Dennett. 2007. *Neuroscience and Philosophy*. New York: Columbia University Press.
- Carruthers, Peter. 2006. *The Architecture of the Mind*. Oxford: Oxford University Press.
- . 1996. *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.
- Chalmers, David. 1997. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
- Churchland, Paul. 2012. *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA: The MIT Press.
- . 2002. "Outer Space and Inner Space: The New Epistemology." *Proceedings and Addresses of the American Philosophical Association* 76 (2): 25–48.
- . 1996. *The Engine of Reason, the Seat of the Soul*. Cambridge, MA: The MIT Press
- . 1989. *A Neurocomputational Perspective*. Cambridge, MA: The MIT Press.
- . 1979. *Scientific Realism and the Plasticity of Mind*. Cambridge, UK: Cambridge University Press.
- Clark, Andy and David Chalmers. 1998. "The extended mind." *Analysis* 58 (1): 7–19.
- Dennett, Daniel. 2007. with Peter Hacker, Maxwell Bennett, and John Searle. *Neuroscience and Philosophy*. New York: Columbia University Press.
- . 1997. *Brainchildren: Essays on Designing Minds*. Cambridge, MA: The MIT Press
- . 1986. *Content and Consciousness*. New York: Routledge.
- Fodor, Jerry. 2000. *The Mind Doesn't Work that Way: The Scope and the Limits of Computational Psychology*. Cambridge, MA: MIT Pres.
- . 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
- . 1981. *RePresentations: Philosophical Essays on the Foundation of Cognitive Science*. Cambridge, MA: The MIT Press.
- . 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press
- Goldman, Alvin and Frederique de Vignemont. 2009. "Is social cognition embodied?" *Trends in Cognitive Science* 13 (4): 154–159.

- Gopnik, Alison and Elizabeth Seiver. 2009. "Reading Minds: How Infants Come to Understand Others." *Zero to Three* 30 (2): 28–32.
- Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
- Hobson, Peter and Jessica Hobson. 2012. "Joint Attention or Joint Engagement? Insights from Autism." In *Joint Attention*, edited by Axel Seemann, 115–136. Cambridge, MA: The MIT Press.
- Hutto, Daniel. 2012. "Elementary Mind Minding, Enactivist-Style." In *Joint Attention*, edited by Axel Seemann, 307–342. Cambridge, MA: The MIT Press.
- Jackson, Frank. 1998. "What Mary Didn't Know." In *Mind, Method and Conditionals*. London: Routledge.
- Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Lakoff, George, and Mark Johnson. 1999. *Philosophy in the Flesh*. New York: Basic Books.
- Leavens, David. 2012. "Joint Attention: Twelve Myths." In *Joint Attention* edited by Axel Seemann, 43–72. Cambridge, MA: The MIT Press.
- . 2005. J.L. Russell and W.D. Hopkins. "Intentionality as measured in the persistence and elaboration of communication by chimpanzees." *Child Development* 76: 291–306.
- . 2004. W.D. Hopkins and R.K. Thomas. "Referential communication by chimpanzees." *Journal of Comparative Psychology* 118: 48–57.
- Noë, Alva. 2010. *Out of Our Heads*. New York: Hill and Wang.
- . 2006. *Action in Perception*. Cambridge, MA: Bradford
- Niedenthal, Paula, Lawrence Barsalou, Piotr Winkielman, Silvia Krauth-Gruber, and François Ric. 2005. "Embodiment in Attitudes, Social Perception, and Emotion." *Personality and Social Psychology Review* 9 (13): 184–211.
- Povinelli, Daniel, J.E. Reux, D.T. Bierschwale, A.D. Allain, and B.B. Simon. 1997. "Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees." *Cognitive Development* 12: 424–461.
- Povinelli, Daniel, K.E. Nelson and S.T. Boysen. 1992. "Comprehension of role reversal in chimpanzees: Evidence of empathy?" *Animal Behavior* 43: 633–640.
- Prinz, Jesse. 2000. "The Duality of Content." *Philosophical Studies* 100 (1): 173–189.
- Rey, Georges. 1992. "Sensational Sentences Switched." *Philosophical Studies* 68: 289–319.
- Rockwell, Teed. 2010. "Extended Cognition and intrinsic properties." *Philosophical Psychology* 23 (6): 741–757.
- . 2008. "Dynamic Empathy: A new formulation for the simulation theory of mind reading." *Cognitive Systems Research* 9: 52–63.

- . 2007. *Neither Brain nor Ghost: A Nondualist Alternative to the Mind-Brain Identity Theory*. Cambridge, MA: The MIT Press.
- . 1995. "Beyond Eliminative Materialism: Some unnoticed implications of Paul Churchland's Pragmatic Pluralism." Unpublished.
- Searle, John. 1992. *The Rediscovery of Mind*. Cambridge, MA: MIT Press.
- . 1990. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the APA* 64 (3).
- . 1984. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press
- . 1983. *Intentionality*. Cambridge, UK: Cambridge University Press.
- . 1980. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 3 (3): 417–424.
- Siegel, Susanna. 2012. *The Contents of Visual Experience*. Oxford: Oxford University Press.

Journal of Cognition and Neuroethics

The Self-Awareness of Reason in Plato

Daniel Bloom

West Texas A&M University

Biography

Daniel Bloom's area of expertise is Ancient Western Philosophy, though his interests in philosophy are very broad. He has a book coming out in 2014 entitled *The Unity of Oneness and Plurality in Plato's Theaetetus*. It addresses the relation between a transcendent principle of being and particular beings. Upcoming projects include an article on the relation between artistry and self-consciousness in Nietzsche's *Gay Science*, and an article on false opinion in Plato's *Theaetetus*. Daniel received his PhD from the University of Georgia in 2012. He will be a professor of philosophy at West Texas A&M University beginning in the Fall of 2014.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Bloom, Daniel. 2014. "The Self-Awareness of Reason in Plato." *Journal of Cognition and Neuroethics* 2 (1): 95–103.

The Self-Awareness of Reason in Plato

Daniel Bloom

Abstract

Socrates is perhaps most famous for two claims: that the life of inquiry (a life dedicated to the use of reason) is the only life worth living, and that he knows that he knows nothing. These two famous claims appear to be starkly at odds with one another. After all, if Socrates really knows nothing then how can he possibly be sure enough to make the rather serious claim that there is only one kind of life worth living? What gives reason such a lofty position if it is not knowledge?

The solution to this apparent problem not only uncovers the heart of Plato's moral theory, but it also sheds light on what it is that sets reason apart from the other faculties. I argue that the fundamental feature of reason relevant to the resolution of the apparent problem is its self-awareness. More specifically, reason's ability to reflect on an idea without internalizing it, or accepting it as true, allows for an investigation into what is or is not good that does not commit the thinker to any particular conclusion. This ability to investigate without accepting, Socrates argues, is what allows us to navigate the world without harming ourselves (and hence is the only life worth living).

However, reason's self-awareness, while it does solve the Socratic problem raised above, leads to a new problem: how can reason (or the reasoning soul) be both the subject and object of its own investigation? Even if reason can be its own object, then won't self-aware reason differ from itself precisely insofar as it is one part subject and one part object, and since the subject would only be aware of the object it wouldn't really be self-aware? The conclusion of the paper will suggest the method Plato uses to navigate this problem, which includes the recognition of the need for intellection, or *nous*.

Keywords

Plato, Inquiry, Ignorance, *Nous*, Reason, *Dianoia*, hypothesis, self-reflexivity, self-awareness, *Meno*, *Apology*

This paper is intended to give a general account of the ground of Socratic virtue, identifying both what it means to be virtuous, and how Plato directs us towards recognizing this. This is available to us, I argue, in spite of the fact that Socrates never gives us a definition of virtue, but rather restricts himself to undermining the accounts of virtue offered by others. For the argument in this paper, I will appeal to the *Meno* and the *Apology*, though I think that the account I am offering holds for all of Plato's dialogues.

Two of Socrates' most famous claims are that the life of inquiry (a life dedicated to the use of reason) is the only life worth living, and that he knows that he knows nothing (Plato 2002, *Apology*, 38a; 22d). These two famous claims appear to be starkly at odds with one another. After all, if Socrates really knows nothing then how can he possibly be

sure enough to make the rather serious claim that there is only one kind of life worth living? What gives reason such a lofty position if it is not knowledge?

Resolving this apparent contradiction in Socrates' reasoning is key to understanding Plato's moral theory, and it also sheds light on what sets reason apart from the other faculties. These are the two primary parts of this paper. First, I will raise and resolve the apparent contradiction that Socrates both knows and does not know the right way to live. I will then argue that this resolution posits a way of acting that is grounded in our ability to reason from hypothesis, and that this ability is, in turn, grounded in reasons self-awareness.

The Apparent Contradiction

One of the definitions of virtue offered in Plato's *Meno* is "to find joy in beautiful things and have power." In response to this definition Socrates presents an argument concluding that all people only desire the good.¹ The central assumption in the argument is that no one desires to harm themselves. Since desiring what is bad is wishing to secure something harmful for oneself, and securing what is harmful for oneself is harming oneself, no one ever does desires what is bad. Yet, it is obvious that people do harm themselves (Anytus and Socrates both suggest the other does so [Plato 2002, *Meno*, 91a–95a]). How can this be? How is it that people ever secure what is harmful for themselves when all they ever seek out is what they think is beneficial. The answer is both obvious and key to understanding Plato's conception of human goodness. Individuals only harm themselves by acting for something that they think is good, but is really bad. Only when acting from a false claim of knowledge (of what is good) can people harm themselves. Thus, according to Socrates' argument, all that is needed to avoid harming oneself is to only act from what one knows (to be good).

Thus, only by acting from knowledge can a person be sure to avoid self-harm. Entailed in this argument is the assumption that every action is a claim about goodness. The pursuing of one thing over another (or the avoidance of one thing over another) is a claim about the value of things. For example, I choose to eat the banana because I think it is both pleasurable and will give me the energy I require to finish my task. I take both pleasure and energy to complete my task to be goods. That is, in choosing to eat the banana I am making a claim about the "goodness" of eating the banana. So too, if I choose to smoke the cigarette I am making a knowledge claim about the cigarette—specifically,

1. The argument identifies beauty and goodness. See: Plato 2002, *Meno*, 77b.

I am claiming the cigarette to be (a) good, either because of the pleasure it gives me, or because I think it makes me look like a character in a gangster film. In both cases (eating the banana or smoking the cigarette) I am asserting the goodness of the action, and in so doing I am making a knowledge claim. What I fail to see in the case of the cigarette is that (most likely) in the long run the good that I attribute to smoking (say pleasure) will be outweighed by the harm (poor health and its attendant pain). Thus, in the case of the cigarette my action was determined by a false assumption of knowledge, i.e., that smoking the cigarette is good. It was thinking I knew what was good when I really did not that lead me to harm myself.

There are really two different kinds of assumptions being made in these actions. The first is that a particular action will lead to a particular end, i.e., that the cigarette will give me pleasure. The second is that pleasure is a good thing to seek. In other words, each action assumes something worth acting for, *and* a means of acquiring that something. In either case, it is easy to see how thinking we know what is good when we really do not know will lead to harm; acting for the wrong end means the direction you are trying to move in is harmful, while using the wrong means to a good end will lead you away from the good end towards something else.

Thus, every action is a twofold assumption of knowledge (of what is good), and the only way to avoid harming oneself is to only act from genuine knowledge. Here we arrive at the problem of Socrates mentioned above. He is a man that claims that all he knows is that he does not know, and yet he still chooses one action over another, often times with remarkable conviction. For example, in the *Apology* Socrates is ready to give his life for his conviction that the life of inquiry is the only life worth living. How can Socrates possibly

choose one action over another, let alone be willing to put his life on the line for such a choice, while knowing that he does not know?²

Inquiry as the highest human good (several versions)

The only knowledge Socrates has is that he does not know. We have just seen that the only way to avoid harming oneself is to act only from knowledge. Since the only knowledge Socrates possesses is the knowledge of his lack of knowledge, it follows that the only way for Socrates to not harm himself is to act from the knowledge of his lack of knowledge. But what action follows from the awareness of the need for, and lack of, knowledge? It can only be seeking knowledge! Inquiry is the only action that cannot be harmful to the person who knows only that he does not know, for it is the only action that follows from knowledge (of a lack of knowledge).

In fact, *only* the person who knows that he does not know is capable of inquiry.³ If knowledge is present there is no need or impetus to seek knowledge. Inquiry is the one activity that can *only* be done by the person who does not know. Every other activity

2. Here is an outline of the argument from the *Meno* (Plato 2002, *Meno*, 77b–78b) that concludes that all (self) harm comes from thinking one knows when they do not know.

To desire beautiful things is to desire good things.

Some people desire good things and some people desire bad things.

A person can either desire bad things believing them to be good or believing/knowing them to be bad.

To desire something is to wish to acquire it for oneself.

Those who desire bad things believing them to be good actually desire good things.

To acquire something bad is to harm oneself.

Thus, to knowingly desire something bad for oneself is to desire to harm oneself.

Therefore, to knowingly desire what is bad is to wish to harm oneself.

No one wishes to harm themselves.

Therefore, no one desires what is bad, neither by thinking it is good nor by thinking it is bad.

Therefore, everyone desires what is good.

Therefore, the only way to harm oneself is to acquire what is bad believing it to be good. Or in other words, the only way to harm oneself is to believe one knows what is good when they really do not know what is good.

3. Edward Halper's excellent article, "A Lesson from the *Meno*," presents Socrates' argument for the life of inquiry in these terms. (See: "A Lesson from the *Meno*," *Gorgia-Menon: Selected Papers from the Seventh Symposium Platonicum*, edited by Michael Erler and Luc Brisson, 234–42. Sankt Augustin, Germany: Academia Verlag, 2007).

is improved (Plato argues that it is made good [Plato 2002, *Meno*, 87d–88d]⁴) by the presence of knowledge. Inquiry, on the other hand, is undermined by the presence of knowledge; it is fundamentally connected to ignorance.

There is something obvious about all of this. Plato suggests that knowledge is the highest good (insofar as it means we will never harm ourselves). As the highest good knowledge is the ground for the worth of everything else (i.e., money is good if one knows how to use it, bravery is good if one knows how to use it, etc. [Ibid.]). Thus, until one has knowledge, the acquiring of knowledge must be the one goal that trumps all others, for all things besides inquiry are only made good through knowledge. Thus, inquiry is the highest human good for as long as we do not know what the highest good is. The danger of assuming the wrong highest good is circumvented by choosing the *seeking* of the highest good as the highest human good.

Here is yet another version of the argument that the life of inquiry is the highest human good. All action is action for a supposed good end. Knowing that I do not know what the good is means that I have no way to act, for I am aware that I have no way of determining what is or is not a good end. However, if I act for the sake of seeking knowledge (i.e., inquiring), I can act for the sake of something that does not commit me to any particular end, for what it commits me to is investigating, which itself is not tied to any conclusion.⁵ To say this in yet another way, the highest good is knowledge, for it is knowledge that ensures all actions are good/beneficial. Without actually knowing what is good, the only way I can act so as to be certain I am not acting in opposition to what is good is to act for the sake of knowing. Inquiry is precisely this acting for the sake of knowledge. Thus humans, insofar as we do not have knowledge, have a kind of surrogate highest good. We cannot act in a way that we know is good (since we do not have knowledge), but we can act in a way that we know is not bad, by seeking after the knowledge that makes everything good. This negative principle, not doing harm (as opposed to doing good), is the ground of Socratic morality.

Inquiry/reason as the method of choice

So, inquiry is the highest human good. What, however, are we to understand by inquiry? Couldn't anything be considered an inquiry into what is good? After all, we

4. In this argument Plato uses multiple terms for “knowledge,” concluding that all things are made beneficial by *phronesis* (often translated as practical wisdom) instead of *episteme* (the typical word for knowledge). Nonetheless, I think for the purposes of the paper the point still holds.

5. This is grounded in the claim that inquiry requires lack of knowledge.

have just recognized that every act entails an assumption of good. Isn't the acting on any assumption, therefore, an investigation into whether or not that assumption of what is good is true or false? In other words, the Socratic knowledge of ignorance would seem to entail an ignorance of what inquiry is in just the same way that it entails ignorance of what the highest good is. If we pick an incorrect method (or means) of inquiry we are just as likely to harm ourselves as if we had chosen an incorrect highest good.

What we need is a way to choose particular actions that does not commit us to an assumption of knowledge. We can find a parallel to this kind of action in thought: we are able to think without committing ourselves to believing what we are thinking. In Plato this type of thought is exemplified by mathematics.⁶ Proofs are based on some initial set of assumptions in the form of axioms and theorems. The conclusions generated from the initial assumptions (i.e., the hypotheses) are true only insofar as the hypotheses are true. This allows us to undertake a meaningful investigation while all the while recognizing that the truth of everything that follows from the initial assumptions is entirely dependent upon the truth of the initial assumptions. In short, we can think without necessarily accepting whatever we think as true.

The question is if there is an analog to this in action? Can we act without assuming that we know the proper way to act, in the same way that we can think without committing ourselves to what is thought? The answer to this question is almost too simple; we act without assuming we know the right way to act by acting from reason instead of desire. Desire, because it lacks the self-awareness of reason, unwaveringly asserts an action as good. Reason, because of its ties to the hypothetical, is able to act in a way that maintains the possibility of recognizing its own error. Or, to put this another way, because reason is able to be aware of its own assumptions it is never unshakably bound to any of them. Thus, every action from reason remains open to its own faults, and hence, the actor is always in a position to recognize that his action opposes inquiry, and to alter it.

In summary, both types of assumptions made in action—assuming an end (or good) towards which the action is directed and assuming a means of achieving that end—can either be made while recognizing their hypothetical character or by falsely claiming knowledge. To falsely claim knowledge is harmful. Thus, if we are to avoid harming ourselves we must act while recognizing the hypothetical character of both the means to the end, and the end itself. The life of inquiry is the end that necessarily recognizes its own lack, for inquiry is only possible with the realization that one lacks knowledge.

6. For examples see Plato 2002, *Meno*, 86e–87c; *Republic*, 509d–511e; and, *Theaetetus*, 148d (where Socrates requests a definition of knowledge along the lines of a previously given mathematical definition).

Rationally choosing a means to the life of inquiry allows us to be ready to abandon an action as soon as we recognize that it is in any way opposed to the life of inquiry.

There are many issues and difficulties that arise out of this account of Plato's Socratic moral theory. For example, reason's ability to hypothesize requires its maintaining a separation from its object. This separation, however, may well preclude the possibility of actually acquiring any knowledge—indeed, for Plato, knowledge seems to require overcoming any distinction between the subject and object, but this very distinction is the one required for hypothetical inquiry. Nonetheless, even if knowledge is never attained, the argument for the life of inquiry still holds, for as I have argued, the value of inquiry is not merely instrumental, but rather, the life of inquiry is the good human life even if that inquiry never leads us to knowledge. And this, I believe, is where the real beauty of Plato's account lies; it shows how reason gives us a moral compass even in the face of its inability to provide us with knowledge.

The need for a higher faculty

We see from the above that reason's ability to maintain distance from its object (i.e., to treat it hypothetically) gives humans a way to act without assuming knowledge, and thus provides us with a way of living that is not harmful to ourselves. The distance between reason and its object, however, also has significant epistemic repercussions. The ability to know, Plato suggests,⁷ is directly connected to our ability to overcome the distinction between the subject and the object. In brief, to know an object is for that object to be present for the subject in the same way that the object is present in itself. Yet, the possibility of reason's self-awareness depends upon the separation between the subject and the object. The thinker cannot fully take on the object while still maintaining the hypothetical character of the investigation.⁸

The issues involved in reason's distinction between the subject and object are difficult to parse. The following is a brief introduction into what the distinction entails for Plato. The subject/object distinction is necessary for inquiry. When a subject grasps an object as distinct from itself the subject must view the object as multiple; merely to grasp the object as distinct from the subject, and as distinct from other objects, requires attributing a multiplicity to the object, for, after all, the distinct object requires characteristics that

7. This is a somewhat controversial claim in Platonic literature. For the sake of this paper I will hypothesize it here.

8. The above assumption regarding knowledge overcoming the distinction between subject and object really amounts to the hypothesizing of the non-hypothetical.

both distinguish it from, and relate it to, other objects. Yet, in order for the grasp of an object to be of *one* object, it must be the case that the plurality necessary for that object to be can somehow be grasped as one thing. This is true of all grasps for Plato. If the grasp of an object is distinct, then that object must have parts that distinguish it from other objects (if something were entirely one it would be identical to anything else entirely one). Yet, to *be* at all, an object must be one. Thus, every grasp is both of the object as one thing, and of the object as a collection of parts. I have suggested that the grasp of the object as a collection of parts is the work of reason. What we now see is that in order for this type of work to take place there must be a higher faculty at work as well, a faculty responsible for the oneness of our grasp. This higher faculty (Plato called it *nous*), because it is responsible for grasping the oneness of an object, must overcome the separation of subject and object.

This distinction between reason⁹ and the higher intellectual faculty (*nous*) is very slippery terrain, in large part because the act of differentiation can only be undertaken by reason, and hence never fully capture *nous*. I will conclude this paper by pointing to an experience in thought that illustrates this distinction. When thinking through a problem one must take the parts being considered and grasp how they fit together. An impasse is the inability to reconcile distinct ideas. The reconciliation comes in the form of an insight that overcomes the difference between the parts being considered. That moment of insight when the impasse is reconciled is the operation of *nous*, and in that act (so this position claims) both the opposition between the parts of the object, and the subject and object, is overcome. The feeling that we experience with an insight (i.e., the feeling connected to the “aha!” moment) is the feeling of oneness between the subject and object. Yet, as soon as we begin to inquire into this new insight we once again fall back into the separation between subject and object required for reasoning, and we are once again engaged in inquiry.

References

Plato. 2002. *Five Dialogues: Euthyphro, Apology, Crito, Meno, Phaedo*. Translated by G.M.A. Grube. Indianapolis: Hackett Publishing.

9. Reason, in this paper, refers mostly to *dianoia*, though Plato uses several different terms in the dialogues in relation to the kinds of judgments I am talking about in this paper.

Journal of Cognition and Neuroethics

Reasoning with and without Reasons: The Effects of Professional Culture and Information Access in Educational and Clinical Settings

Barry Saferstein

California State University San Marcos

Biography

Barry Saferstein is a cognitive sociologist, who applies discourse analysis and ethnography to study the interrelationship of understanding, interaction, and authority in organizational settings. His recent research examines clinical consultations, explaining the effects of communication patterns, information resources, and professional culture on patients' understandings of their medical conditions and treatment options. He has also studied explanations and understandings of genetics in schools and science centres. Prof. Saferstein examines how professional authority affects the development of understandings, explaining how unintended strategies that limit understanding are embedded within explicit strategies for comprehension of medical conditions and scientific concepts. His research has also included studies of agenda setting activities and the interactional construction of ideology in television production settings in the United Kingdom and the United States. He is a Professor in the Communication Department at California State University San Marcos.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Saferstein, Barry. 2014. "Reasoning with and without Reasons: The Effects of Professional Culture and Information Access in Educational and Clinical Settings." *Journal of Cognition and Neuroethics* 2 (1): 105–125.

Reasoning with and without Reasons: The Effects of Professional Culture and Information Access in Educational and Clinical Settings

Barry Saferstein

Abstract

This study examines routine interpretation activities in commonplace settings, which integrate environmental information resources and cultural conventions as components of reasoning and reason. It expands on research about coherence based reasoning, mental models, and cognitive sociology. Analysis of recorded genetics learning activities and interventional radiology consultations shows how the operations of reasoning and reason develop through the interpretation activities of producing brief chains of information, process narratives. What counts as a reason relevant to a particular setting is often related to the ways that people cope with an absence of information. Modus tollens problems, the traditional Mendelian genetics curriculum, and conventional medical consultations include forms of expression that restrict information and lead to the interpretive contingencies, which complicate reasoning and understanding. The interpretation activities related to the restriction of information add contingencies that complicate reasoning. Those interpretive contingencies include the activities of seeking, assessing, culling, and applying information. The genetics education data feature the recognition and application of placeholders, grey boxes, to reduce the interpretive contingencies related to restricted information. Grey boxes are meta-information, marking the suspended interpretation activities of searching for more substantive information. They function syntactically by contributing to process narratives that are coherent in relation to particular localized information resources and interpretation activities.

Examples of recorded data from two settings explicate these findings. One setting features high school biology students learning to use the nomenclature and computational devices for explaining the genetic inheritance of traits. The genetics examples show how curricular conventions of biology teachers' professional culture restrict information resources and guide students toward depending on grey boxes to link available information resources as acceptable forms of reasoning. The clinical data further explicate how information resources and interpretation activities become part of reasoning, and constitute the type of reason that is useful in a particular situation. Recordings of an interventional radiology consultation and a post-consultation discussion show how a patient develops an understanding of a medical procedure and the reasoning related to deciding on medical treatment. The patient's recall of reasons and reasoning correlates with particular actions and information resources in the consultation setting. Both the restricted reasoning of the genetics curriculum and the expansive reasoning of the radiology consultations show how interpretation activities and information resources operate as part of the reasons and reasoning that count as particular forms of reason.

Keywords

Reasoning, understanding, professional culture, coherence based reasoning, mental models, modus tollens, process narratives, grey boxes, patient-practitioner interaction, Mendelian genetics

Reason and reasoning are often treated as mental processing. However, analysis of routine interpretation activities in commonplace settings shows that environmental information resources and cultural conventions function as components of reasoning and reason. Reason, reasons, and reasoning often reflect the contingencies of seeking and interpreting information to develop coherent strings of information, process narratives, that are key to understanding and recall. Process narratives are short information chains that link objects, actions, and events, expressing sequential, causal, or temporal relationships (Saferstein 2007; 2014). They distill information and concepts into a memorable form. Process narratives do not necessarily link information as logical propositions. They may involve fuzzy relationships such as categorical associations, shared settings, and particular time frames.

Interpretive Contingencies and Grey Boxes

The contingencies that affect reasoning include the activities of interpreting the variables—i.e., of interpreting, organizing, and seeking information in order to develop understandings. Interpretive contingencies also include the effort of simultaneously interpreting information while dealing with organizational and cultural constraints, such as time limits and authority, expressed through patterns of interaction.

What counts as a reason is often related to the ways that people cope with an absence of information—e.g., by recognizing that information is missing and searching for it. However, such searching is often constrained by the resource environment and communication patterns. When that occurs, people develop process narratives by applying placeholders, grey boxes, to mark missing information and link the available information (Saferstein 2007; 2014). Grey boxes are meta-information, marking the suspended interpretation activities of searching for more substantive information. As reasons, they function syntactically by contributing to process narratives that are coherent only in relation to particular localized information resources and interpretation activities. Grey boxes often facilitate reasoning by eliminating the contingencies related to searching for missing information. Locally and culturally specific grey boxes are the basis of what seem to be different modes of reasoning.

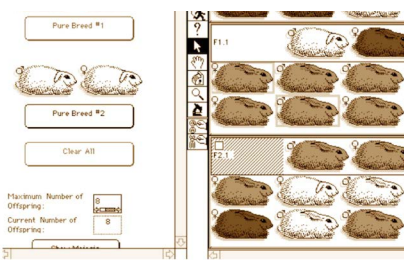
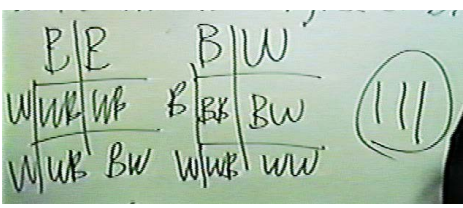
For example, the professional culture of biology teachers and clinicians has been constituted and sustained through the patterns of communication that emphasize certain information and discourage students and patients from seeking other information (Fisher and Groce 1990; Maseide 1991, 2007; Saferstein 2007, 2014). Consequently, what counts as a reason in educational and clinical settings often is related to the absence of information

and the way that the participants in those settings cope with that absence—i.e., not just by recognizing missing information or searching for it, but by applying particular types of grey boxes to mark missing information, and then linking the information that is present by means of those grey boxes.

Reasoning without Reasons

Reasoning without reasons is a feature of certain organizational activities and settings. In the following examples, genetic inheritance learning activities restrict students' use of information about the cellular and molecular processes by which genes affect traits. What counts as reason and reasoning in the conventional Mendelian curriculum is the use of nomenclature and tabulation/computation devices as grey boxes for the missing information. The reasons that contribute to the linking of information as process narratives that explain specific cases of trait inheritance include words and nomenclature that conflate qualities of traits and the biochemical operations of genes. They omit information about the cellular processes by which genes affect the development of traits.

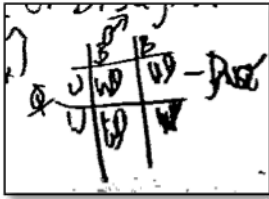
Figure 1: Visual and Linguistic Grey Boxes for Cellular Processes

1	A. Computer Simulation's Representation of Codominant Trait Inheritance	B. Acceptable Punnett Square Representation of Codominant Trait Inheritance
		
2	Acceptable Mendelian Linguistic Representation of Codominant Trait Inheritance (Codominance Test Transcript, Lines 121–123)	
	<p>Student B21:</p> <p>Because in the first line you had two pure breeding rabbits (????) two whites, right?</p> <p>These two genes . you cross bred . them both. All of them became beige . alright . . Alright.</p> <p>And then in the second one there's two genes. That means the two genes they have in them are brown and white. Right?</p> <p>Then when you cross like this one you have two browns a brown and white, white and brown, and a white and white.</p> <p>This is the reason that we came up with three different colors. (??) these two made a beige . and that's one . and we had a darker color. That's two. Then we had a white . three. That's why we got the three colors.</p>	

The following examples occurred during a verbal testing session in a high school biology class.¹ The teacher (T1) asked for an explanation of codominance, which the students had studied by using a computer simulation showing the effects of breeding a rabbit having brown fur and a rabbit having white fur (Figure 1, Row 1, Image A). The rabbits' offspring had beige fur. Breeding the beige rabbits produced a heterogeneous generation of offspring respectively having white, brown, or beige fur. An acceptable Mendelian explanation was presented by a student, B21 (Figure 1). It featured the standard Mendelian grey box, the Punnett square (Figure 1, Row 1, Image B), which tabulates nomenclature representing both genes and traits in order to compute the probable outcomes of breeding individuals having particular sets of genes and traits. Neither the Punnett square nor the linguistic description of its application involves reasoning or reasons related to the cellular genetic processes by which genes affect trait development. Those processes are grey-boxed by the Punnett square and by linguistic descriptions such as, "became beige" and "these two made a beige" (Figure 1, Row 2).

The acceptable Mendelian explanation of codominance features two Punnett squares (Figure 1, Row 1, Image B). The first counts as an explanation of the first generation beige offspring (labeled F1.1 in the simulation [Figure 1, Row 1, Image A]). The second Punnett square counts as an explanation of the heterogeneous second generation (F2.1). The discussion here focuses on an explanation of codominance offered by another student, B23. He began by presenting a verbal explanation. The teacher asked him to clarify it by using Punnett squares (Figure 2).

1. The discussion of genetics learning activities is based on studies in high school biology classrooms and science education centers over a period of 15 years. The classroom data discussed here were recorded in a California High School biology class, as part of the Community of Explorers Project (National Science Foundation RED-9154815).

Figure 2: B23's Attempts to Explain Codominance with Punnett Squares

A



B

After B23 completed one acceptable Punnett square (Figure 2, image A) related to the mating of a brown-furred and white-furred rabbit to produce beige-furred offspring, he did not immediately draw a second Punnett square to explain how the mating of two beige rabbits would produce beige, brown, and white offspring. Instead, he presented a verbal explanation emphasizing the combining of genes to produce the beige rabbits. The teacher then urged him to draw a second Punnett square based on the offspring indexed by the cells of the first Punnett square, i.e., beige rabbits characterized as WB.

However, B23 had not developed an understanding of how to use a Punnett square as a grey box for the missing cellular process information. He proceeded to set up the second Punnett square differently than the first. The teacher did not accept B23's erroneous second Punnett square (Figure 2, Image B) as an explanation of codominance. Nor did he recognize B23's verbal explanation of codominance as an example of reason and reasoning about codominant trait inheritance. However, examining B23's verbal explanation reveals reasoning that would count as reason in regard to the cellular processes involved in codominance.

Rather than presenting the conventional tabular and computational explanations of operating a Punnett square, B23's reasoning consisted of descriptive comments that correlated with computer simulation's images of generations of rabbits, such as:

"these two they met in the midpoint codominant" (line 59)

"They share two dominant--the dominant brown gene and the dominant white gene come together and they form uh somewhere at midpoint like . beige." (line 89)

"this brown and the white both made the beige" (line 93)

“these two genes white--white and brown, are in the beige ones are in the beige rabbit” (line 97)

“you would have--you would have the two dominant genes and they would come out in the gener--in the F2 generation” (line 101)

B23 linked information as process narratives, imputing a direct causal connection between the presence of certain genes in the bodies of the rabbits and the fur color of those rabbits. He accurately suggested that mating two of the beige rabbits, each having a gene for brown fur and a gene for white fur, would result in a set of their offspring having the two dominant genes. He recognizes that there was no specific gene for beige.

Utterances referring to dominant genes that “come together” to “form . . . somewhere in the midpoint” and “made the beige” could index cellular genetic processes, and function as a display of reason based on the reasoning related to the production and effects of enzymes and proteins on the development of rabbits’ fur color. However, due to an absence of such cellular process information B23 relies on ‘form’ and ‘made’ as grey boxes in his process narratives.

As a description, B23’s process narratives are not without reason, reasoning, or reasons. The explanation is reasonable in the context of reasoning based on genes having properties that affect the development of traits. It combines observation of the computer simulation images (Figure 1, Row 1, Image A) with some of the Mendelian terminology (e.g., ‘dominant’, ‘F2 generation’). However, B23’s reasoning was not recognized as reasoning or as a display of reason by the teacher, because it did not feature the Mendelian grey box, a Punnett square.

The teacher’s response shows that his interpretive frame of reference emphasizes the customary linguistic and graphical representations that function as reasons within the professional culture related to teaching Mendelian genetics. In that context, he does not recognize the reasons expressed by B23 as components of reasoning. The teacher says, “I’m trying to understand. You’re throwing words at me. I can’t understand” (line 96). The teacher’s response displays the effects of professional culture on recognizing reason.

Even if B23 were to some extent improvising an explanation, the process narratives he expressed display the reasoning indicating that the genes come together to affect traits. As much as it might seem that he is just “throwing words” at the teacher, they are not just random words. He links certain words representing combinations of types of genes to words that represent particular fur colors as a model of causation. There is reason in B23’s utterances.

B23's reasons and reasoning not only reflect the absence of cellular process information and the availability of the computer simulation's visual information, but also his interpretation activities during his task group's prior work with the computer simulation. At that time, he had monitored how other task groups developed explanations of the codominance simulation and reported back to his group. Consequently, he did not fully participate in the interpretation activities by which the other student, B21, developed his acceptable Mendelian explanation of codominance (Figure 1, Row 2). B23's interpretation activities related to available information resources and his searching for additional information shaped his reasons and reasoning.

Missing Reasons: The Interpretive Contingencies of Modus Tollens

In the conventional curriculum and pedagogy for Mendelian genetics, the professional culture of biology teachers has featured communication patterns that discourage students searching for cellular process information (Saferstein 2014). This limits the information that students can use to link as reasons about types of genes and types of traits in order to explain trait inheritance. Such patterns of explanatory interaction between students and teachers create a 'not cellular processes' contingency. The teachers apply particular types of grey boxes as placeholders for the missing information. These include linguistic terms (e.g., 'recessive'), nomenclature conflating genes and traits (e.g., WB), and Punnett squares (Saferstein and Sarangi 2010). However, none of these provide reasons that would contribute to reasoning explaining the cellular and molecular processes by which genes function within the body to produce traits.

Such restriction of information also occurs in formal logic problems. For example, studies of reasoning have noted the difficulty in solving modus tollens logic problems (i.e., If p then q . Not q . Therefore, not p) (cf., D'Andrade 1989; Johnson-Laird 1983; Johnson-Laird and Byrne 2002; Johnson-Laird et al. 1972; Wason 1968). The difficulty some people have in solving modus tollens problems and the difficulty some students have in accepting the Mendelian grey boxes are analogous. In the 'not q ' expression of a variable, negation can be problematic due to its lack of specificity, a condition that routinely triggers the contingencies of searching for relevant information.

In regard to effects on the interpretation activities that produce reasoning, there is a similarity between the genetics curriculum's emphasis on *not applying cellular process information* and the premise of modus tollens problems of *not applying contextual information regarding some real-world relationship between the variables*. In both cases, the difficulties in producing acceptable reasoning do not result from the expressed logical

contingency between variables. Rather, they result from the contingency of interpreting the variables or a relationship between the variables in a way that transcends the grounds stated by the problem.

When explanations or problems do not provide adequate information for developing coherent process narratives that would contribute to understandings or solutions, people face the interpretive contingencies of searching for missing information (Saferstein 2007; 2014). In that context, it is difficult not to search for reasons in order to engage in reasoning—i.e., reasons in the form of contextual relationships based on experiential recall or inference. This often involves comparing available information with previously developed process narratives in regard to categorical distinctions such as typology, situation, and setting. In the cases of Mendelian explanations of trait inheritance and modus tollens problems, such comparison involves considering information that is not expressed by the problem as stated (Saferstein 2014; D'Andrade 1989). The comparison contributes to a frame of reference that supports searching for relevant information, rather than applying a restrictive puzzle-solving frame of reference that emphasizes literal use of the stated variables.

For example, contemporary high school students have encountered information inside and outside of the biology classroom about genomics and the role of DNA in the development of organisms (Saferstein 2014). Yet, the conventional Mendelian curriculum requires that the students do not apply that information when trying to explain trait inheritance. Thus, like some formulations of modus tollens problems, the students are presented with information ('the inheritance of genes affects the expression of traits') that relates to process narratives they have already created (e.g., 'DNA and RNA in cells affect traits'). When biology students initially encounter the emphasis on the abstract Mendelian grey boxes and the de-emphasis of cellular processes, some of them continue to mention commonplace topics such as families, breeding of animals, and DNA, in attempts to form explanatory process narratives (Saferstein 2014). In the preceding example, B23's explanation of codominance suggests the operation of cellular processes (the blending of genes), rather than applying the computational grey box of a Punnett square. However, applying information or a frame of reference related to biochemical cellular processes results in an unacceptable answer based on unacceptable reasoning.

As biology teachers restrict discussion of cellular processes during genetic inheritance learning activities they guide students toward reducing the interpretive contingencies related to the 'not cellular processes' restriction by accepting and applying the conventional Mendelian grey boxes. Rather than repeatedly confronting the interpretive contingences presented by constraints on seeking missing information, students ultimately reduce

those contingencies by relating nomenclature and Punnett squares to trait inheritance (Saferstein 2014). For the students, Mendelian grey boxes become reasons. For the teachers, correct manipulation of those grey boxes functions as a student's display of reasoning and understanding. The encompassing framework of reason is shaped by interpretation activities that include the particular communication patterns, which constitute the professional culture of biology teachers. Such reason involves the use of the grey boxes in particular ways. The reasoning is the linking of the grey boxes with other information to create the process narratives.

Modus tollens problems, the traditional Mendelian genetics curriculum, and conventional medical consultations include forms of expression that restrict information and lead to the interpretive contingencies, which complicate reasoning and understanding. Johnson-Laird (1972) and D'Andrade (1989) conducted experiments finding that the use of commonplace objects and activities improved the understanding and solving of modus tollens problems. D'Andrade concluded that such changes resulted from replacing a contentless sense of contingency with a contentful sense of contingency. However, examining the production of process narratives provides another approach to understanding why abstractly expressed modus tollens problems are more difficult to solve than problems expressed in terms of common experience. It is not a particular *sense* of contingency, but the interpretation activities of contending with contingencies that inhibit or lead to reasoned understandings.

Information Components That Become Reasons: Interpreting Linguistic, Pictorial, and Gestural Information

Like the Mendelian genetics students, patients in conventional clinical consultations, often experience limited forms of explanatory information and restricted opportunities to seek such information (cf. Fisher 1986, 1993; Fisher and Groce 1990; Frankel 1990; Maseide 1991, 2007; Mishler 1984). The resulting interpretive contingencies lead to situations that are familiar to many patients: i.e., not knowing what questions to ask of practitioners during a consultation, thinking of the questions after the consultation has concluded, and not recalling or understanding medical information discussed during a consultation (cf. Entwistle et al. 2006; Price et al. 2006; Skea et al. 2004). As a result, patients often do not develop reasons relevant to reasoning about the nature of medical conditions and the effects of treatment options. This creates 'you had to be there' moments, when a patient has difficulty reconstructing a practitioner's explanation of a medical procedure, which had seemed clear during the consultation. Missing information or constraints on

searching for information change the production of the process narratives central to reasoning from the linking of substantive information to the use of culturally specific grey boxes that are only syntactically meaningful in regard to a particular setting.

The following example provides a contrast to conventional clinical consultations and the Mendelian genetics learning activities. During an interventional radiology consultation, a patient, P1, and nurse, N1, verbally and gesturally respond to each other and to a series of images displayed on a computer in order to explain uterine fibroid tumors, their symptoms, and treatment options, which included uterine fibroid embolization (UFE). The consultation involves interpretation activities that avoid the contingencies related to missing information and support the patient's creation of recallable process narratives.²

The process narratives that P1 recalled four days later, during a recorded telephone discussion with the author, linked and condensed information expressed by the verbal and visual information she encountered at the consultation. For example, during the telephone discussion, P1 expressed process narratives concerning what she should do if she recognized symptoms of a dangerous complication, sloughed necrotic fibroid (SNF), consequent to a uterine fibroid embolization procedure. She presented an overview of a nurse's explanation, which had linked the term, 'sloughed necrotic fibroid', to symptoms and treatment. The reasons and reasoning constituting those process narratives correlate with the moments during the radiology consultation when the patient completed the nurse's utterances, voiced comprehension or participation, and turned her head toward images and gestures that related to the nurse's verbal explanation.

During the telephone discussion, P1 expressed process narratives concerning the possibility of needing a D & C procedure (dilation and curettage) due to a sloughed necrotic fibroid (SNF) that could lead to a uterine infection:

some of the . things that I can . are uh they want me to be aware of after the procedure. Right that-that I may need a D & C--I mean there--I forget what the percentage . uh what'd she say? . Two or three percent I think (Recorded P1 Telephone Discussion, lines 1, 3, 5)

afterwards when the fibroids are . you know if they die or whatever eh-eh uh I might have . there's a possibility of a . hu uterine infection. I

2. This data is part of an ongoing study of the effects of discussing images on patterns of patient-practitioner communication that affect patients' understandings of medical information about medical conditions and their treatment.




have to be aware of the um . . . yeah of the symptoms of that (Recorded P1 Telephone Discussion, lines 13, 15, 17)

P1's reasoning combines a cause (*for approximately 2 or 3 percent of UFE patients, a type of fibroid tumor may 'die' as a result of UFE*) and an effect (*the 'dead' tumor causes a uterine infection*) with actions (*patient awareness of the potential symptoms of infection, treatment with a D & C procedure*). The reasons and reasoning P1 expressed during the telephone discussion correlated with particular talk, gestures, and images presented during the radiology consultation.

Turning Actions into Reasons

During the radiology consultation, P1's interpretation of the nurse's verbal explanations occurred in the context of pointing and looking at the images on the computer screen as well as displays of attention to the nurse's utterances. For example, the information about a fibroid falling from the lining of the uterus into the uterine cavity included gestures by the nurse as she explained the screen text concerning SNF (Figure 3, Row B). As she said, "a fibroid is closer to the lining of the uterus and it actually falls . . . into the center cavity of the uterus," the nurse made a fist with her right hand, illustrating the fibroid tumor, and moved it down to her left cupped left hand, illustrating the tumor falling into the uterine cavity (Figure 3, Row B). P1 looked at the nurse gesturing, and then completed the nurse's explanation of the fibroid falling into the uterus, saying, "and tries to come out through the uterus." The nurse's gestures also functioned as reasons contributing to the reasoning expressed by the process narratives that P1 recalled during the post-consultation telephone interviews.

Figure 3: Modes of Representation for Sloughed Necrotic Fibroid³

<p>A</p>	 <p>1</p>	<p>UFE Complications</p> <ul style="list-style-type: none"> • Complications of angiography 1:1000 • Transcervical expulsion fibroid tissue (~5%) • Infection-hysterectomy 1:250 • Premature ovarian failure (menopause) • Mortality: 1:4000 (?) <p>2</p>	<p>Postprocedure care</p> <p>Impacted tissue may require GYN intervention of a hysteroscopic D & C, not necessarily hysterectomy.</p> <p>Patients and other MDs will NOT recognize this complication!</p> <p>3</p>
<p>B</p>	<div style="display: flex; justify-content: space-between;"> <div style="width: 30%;">  </div> <div style="width: 65%;">  </div> </div> <p>505. N1: this is actually our main concern. And it happens in less than 5 percent of the patients and it's called sloughing a necrotic fibroid. In other words a fibroid is closer to the lining . of the uterus and it actually falls into the lining of—or into the center cavity of the uterus.</p> <p>506. P1: and tries to come out through the cervix</p> <p>507. N1: Well um eh . of those five percent, 95 percent of them just get passed in your cycle.</p> <p>508. P1: Okay.</p>		

3. Pictures of the consultation have been captured from the video data and faces have been blurred in order to protect the participants' privacy.

Turning Images into Reasons

During the post-consultation telephone discussion, P1 recalled a picture that had been presented 7 minutes and 50 seconds prior to the nurse's explanation of SNF:

- 2.2.17 Researcher: You mentioned uh the possibility of infection afterwards as something that you hadn't uh you
- 2.2.17 P1: I hadn't yeah I didn't I I knew nothing about yeah.
- 2.2.19 Researcher: Do you remember any of the um discussion or the the PowerPoint that related to . to that—that sticks in your mind?
- 2.2.19 P1: Um . okay. Yyeah .. the th-um there was a diagram of a uterus . and fibroids um and I guess there was a picture of one right over the . at the cervix or something.

The picture showed a large fibroid tumor within the uterine cavity, which blocked the cervix. This visual information was consistent with N1's later gesture of her fist falling into her cupped hand as she said, "a fibroid is closer to the lining of the uterus and it actually falls into the lining of—or into the center cavity of the uterus."

Just prior to the nurse's verbal and gestural information about SNF, the computer screen had displayed following textual information:

- Infection-hysterectomy 1:250 (Figure 3, Row A, Image 2)
- Impacted tissue may require GYN intervention of a hysteroscopic D & C, not necessarily hysterectomy. (Figure 3, Row A, Image 3)

While the computer screen displayed the list concerning SNF (Figure 3, Row B), the nurse's explanation presented eleven pieces of information that the patient could interpret as reasons and link as process narratives (bold text) (Figure 3, Row B, lines 505, 507, 509):

- "This is actually **our main concern**
- And it happens **in less than 5 percent of the patients**
- And **it's called sloughing a necrotic fibroid.**
- In other words **a fibroid is closer to the lining of the uterus and it actually falls** into the lining of—or **into the center cavity of the uterus**
- Well um eh **of those five percent, 95 percent** of them [sloughed necrotic fibroids] **just get passed in your cycle**
- **That five percent that's left over if they get stuck there and they can't go out, then it's gonna set you up for infection.**
- So what da we need to do? **You need to call us.**

- **We need to do an MRI right away—bring you in the hospital.**
- **I.V. antibiotics.**
- Do an MRI. **See what's going on,**
- **And we may ask your gynecologist to do a D & C."**

During the telephone discussion four days after the consultation, P1 linked some of the verbal and visual information encountered during the consultation as reasons for SNF: a 'dead' fibroid blocks the cervix, the fibroid is trapped in the uterus, infection develops, treatment with D & C. Her process narratives correlate with certain interpretation activities and information resources of the radiology consultation. The culling and organizing of the initial set of information constitutes P1's reasoning.

Discussion of images at the radiology consultation provided patients with options for interpreting and linking information—i.e., opportunities for constructing the reasons and reasoning that contribute to reasonable explanations of medical conditions and treatments. Such options not only included multiple types of information, but also a communication pattern by which the patients developed frames of reference supporting unanticipated linkages of information. These interpretation activities avoid the modus tollens-like interpretive contingencies that entail the use of grey boxes to mark missing information. Both the restricted, grey-boxed reason of the genetics curriculum and the expansive reason of the radiology consultations show how interpretation activities related to the information resources of a setting produce the reasons and reasoning that count as particular forms of reason.

Process Narratives and Mental Models

The preceding examples explicate the differences between mental model research and the process narrative approach to understanding/recall. As relevant interpretive contingencies, the mental model approach emphasizes contingent relationships among available variables (cf. Bauer and Johnson-Laird 1993; Johnson-Laird 2002). For example, Johnson-Laird compares the use of visual representations of variables and conditions developed by Peirce with mental models in order to explain fundamental operations of reasoning:

The fundamental operations of reasoning based on mental models are insertion (the addition of entities, properties, or relations to models), and deletion (the elimination of models when they are combined with other, inconsistent, models). In the case of reasoning based

on quantifiers, the theory also proposes that individuals search for alternative models. (Johnson-Laird 2002, 91)

That approach emphasizes the mental organization of interpretive artifacts rather than the processes of interpretation. It does not consider the contingencies of creating, finding, and organizing alternative models as part of the reasoning. Information constituting the “entities, properties, or relations to models” mentioned by Johnson-Laird must be interpreted in order to function in such capacities. However, as the preceding clinical example shows, such interpretation is not separate from the resulting “operations of reasoning.” Insertion and deletion of interpreted information are not just mental operations, but also involve a complex of interrelated interpretation activities, information resources, and sociocultural constraints of settings. Johnson-Laird’s “elimination of models when they are combined with other, inconsistent, models” occurs during the interpretation of information as well as during the reasoning involving the interpretations (Saferstein 2007; 2014). Examining recorded interpretation activities, such as P1’s medical consultation, and their effects on understanding shows people interpreting information prior to and while inserting and deleting it. This reveals the interrelated development of reason, reasoning, and reasons. That interrelationship is due to the function of environmental information resources as components of mind in regard to organizing and recalling information as understandings.

Mind and cognition do not just rely on resources in settings as opportunities for perception to obtain information that is then subjected to mental operations in order to produce reason or meaning. Rather, environmental information resources, sometimes including social interactions, are also part of the operations of reason. They function as placeholders when information related to producing coherent linkages is absent, i.e., as external ‘storage’ that relieves the burden on memory during reasoning, and as components of recall afterward (Saferstein 2014).

Reconfiguring Cognitive Systems and Reasoning

Analyzing the creation and use of process narratives and grey boxes reveals the interpretation activities that constitute a cognitive system of reason, reasoning, and reasons. This elucidates a central concern of coherence-based reasoning research, the idea that a cognitive system *imposes* coherence on decision tasks (Simon 2004: 517, 522–523; cf. Holyoak and Simon 1999). Moreover, by considering the effects of material resources and professional cultures on reason and reasoning, such analysis changes the conceptualization of ‘cognitive system’ from a focus on mental calculations, logical

choices between predetermined decision options, or mental models. It explicates how a cognitive system includes the resources of the setting and the cultural conventions that support the use of the resources in the setting (cf. Simon 2004: 517, 522–523). Such analysis expands and elucidates the simultaneous development of interpretations of information and frames of reference discussed in studies of coherence based reasoning (e.g., Holyoak and Simon 1999; Simon 2004; Simon et al. 2004; Pennington and Hastie 1992; Thagard 1989), as well as the reflexive relationship of interpretation activities and social organization discussed in cognitive sociological research (Cicourel 1973; Saferstein 1994; 2007; 2010). It also clarifies abstract categories, such as retrieval, segment linkage, and distillation of memories, which focus on mental operations (cf. Schank and Abelson 1995, 31–33, 71–72).

The genetics examples show that the teacher's restricting of information led the biology students toward depending on grey boxes to link available information resources of the setting in order to create coherent process narratives. The external information resources of the setting, including the Mendelian grey boxes and the interpretive interaction, were as much a part of the reasons and reasoning as the mental operations. This was explicated in the second example when the patient's recall of reasons and reasoning correlated with the actions and information resources in the consultation setting. The access to relevant information resources in the clinical example shows how interpretation activities, including the culling and recognition of information to produce coherent recallable process narratives that function as reasons, become part of reasoning, and constitute the type of reason deemed useful in a particular situation.

Such findings emphasize that accurate models of reasoning, cognition, and memory would incorporate interpretation activities that feature information resources. Recognizing such material components of reason is particularly relevant to understanding the relationship between repeated routine reasoning related to commonplace settings and the neuroplastic development of the brain's organization of information as reason.

Reasoning is the set of interpretation activities that develop coherence by emphasizing or excluding the particular information to be linked. Reasons consist of both the frames of reference resulting from those interpretation activities and the process narratives produced by linking the selected pieces of information. These are concurrently developed. However, when professional or organizational culture introduces barriers to these activities, interpretive contingencies trouble reasoning and understanding. Pragmatically, reason is a result of reasoning activities; not a template controlling them.

References

- Bauer, Malcolm I. and P.N. Johnson-Laird. 1993. "How diagrams can improve reasoning." *Psychological Science* 4 (6): 372–378.
- Bobrow, Danny G. and Donald A. Norman. 1975. "Some principles of memory schemata." In *Representation and Understanding: Studies in Cognitive Science*, edited by D. G. Bobrow and A. M. Collins, 131–149. New York: Academic Press.
- Cicourel, Aaron V. 1973. *Cognitive Sociology: Language and Meaning in Social Interaction*. Harmondsworth: Penguin.
- D'Andrade, Roy G. 1989. Culturally based reasoning. In *Cognition and Social Worlds*, edited by A. Gellatly, D. Rogers and J. A. Sloboda, 132–143. Oxford: Oxford University Press.
- Entwistle Vikki, Brian Williams, Zoe Skea, Graeme MacLennan, Siladitya Bhattacharya. 2006. "Which surgical decisions should patients participate in and how? Reflections on women's recollections of discussions about variants of hysterectomy." *Social Science & Medicine* 62: 499–509.
- Fisher, Sue. 1986. *In The Patient's Best Interest*. Piscataway, NJ: Rutgers University Press.
- Fisher, Sue. 1993. "Doctor talk/patient talk: How treatment decisions are negotiated in doctor-patient communication." In *The Social Organization of Doctor-Patient Communication* (2nd ed), edited by S. Fisher and A. Todd, 161–182. Norwood, NJ: Ablex.
- Fisher, Sue and S. B. Groce. 1990. "Accounting practices in medical interviews." *Language in Society* 19: 225–50.
- Frankel, Richard. 1990. "Talking in interviews: a dispreference for patient-initiated questions in physician-patient encounters." In *Interactional Competence*, edited by George Psathas, 231–262. Lanham, MD: University Press of America.
- Holyoak, Keith J., & Simon, Dan. 1999. "Bidirectional reasoning in decision making by constraint satisfaction." *Journal of Experimental Psychology: General* 128 (1): 3–31.
- Johnson-Laird, P. N. 1983. *Mental models*. Cambridge, MA, Harvard University Press.
- Johnson-Laird, P.N. 2002. "Peirce, logic diagrams, and the elementary operations of reasoning." *Thinking and Reasoning* 8 (1): 69–95.
- Johnson-Laird, P. N. 2006. "Models and heterogeneous reasoning." *Journal of Experimental & Theoretical Artificial Intelligence* 18 (2): 121–148.
- Johnson-Laird, P. N. and Byrne, Ruth M. J. 2002. "Conditionals: A Theory of Meaning, Pragmatics, and Inference." *Psychological Review* 109 (4): 646–678.
- Johnson-Laird, P. N., Paolo Legrenzi, and M. S. Legrenzi. 1972. "Reasoning and a sense of reality." *British Journal of Psychology* 63: 395–400.

- Måseide, Per. 1991. "Possibly abusive, often benign, and always necessary: On power and domination in medical practice." *Sociology of Health and Illness* 13 (4): 545–561.
- Måseide, Per. 2007. "Discourses of Collaborative Medical Work." *Text and Talk* 27 (5/6): 611–632.
- Mishler, Elliot G. 1984. *The Discourse of Medicine: Dialectics of Medical Interviews*. Norwood, NJ: Ablex.
- Pennington, Nancy and R. Hastie. 1992. "Explaining the evidence: Testing the story model for juror decision making." *Journal of Personality and Social Psychology* 62: 189–206.
- Price, J., G. Farmer, J. Harris, T. Hope, S. Kennedy, and R. Mayou. 2006. "Attitudes of women with chronic pelvic pain to the gynaecological consultation: A qualitative study." *British Journal of Obstetrics & Gynaecology* 113 (4): 446–452.
- Saferstein, Barry. 1994. Interaction and ideology at work: A case of constructing and constraining television violence. *Social Problems* 41 (2): 316–45.
- Saferstein, Barry. 2007. "Process narratives, grey boxes, and discourse frameworks: Cognition, interaction, and constraint in understanding genetics and medicine." *The European Journal of Social Theory* 10 (3): 424–447.
- Saferstein, Barry. 2010. Cognitive Sociology. In *Society and Language Use*, edited by Jürgen Jaspers, Jan-Ola Östman, and Jef Verschueren, 113–126. Amsterdam, The Netherlands: John Benjamins.
- Saferstein, Barry. 2014. *Understanding and Interaction in Clinical and Educational Settings*. Sheffield, U.K.: Equinox.
- Saferstein, Barry, and Srikant Sarangi. 2010. "Mediating modes of representation in understanding science: The case of genetic inheritance." In *Exploring Semiotic Remediation as Discourse Practice*, edited by Paul A. Prior and Julie A. Hengst, 156–183. Houndmills, Basingstoke, U.K.: Palgrave Macmillan.
- Schank, Roger C. and Robert P. Abelson. 1995. "Knowledge and memory: The real story." In *Knowledge and Memory: The Real Story*, edited by R. S. Wyer Jr., 1–85. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Simon, Dan 2004. "A third view of the black box: Cognitive coherence in legal decision making." *University of Chicago Law Review* 71: 511–586. USC Public Policy Research Paper No. 04-10. Los Angeles: University of Southern California Law School.
- Simon, Dan, C. J. Snow, and S. J. Read. 2004. "The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction." *Journal of Personality and Social Psychology* 86 (6): 814–837.

- Skea, Zoe, V. Harry, Siladitya Bhattacharya, Vikki Entwistle, Brian Williams, Graeme MacLennan, A. Templeton. 2004. "Women's perceptions of decision-making about hysterectomy." *British Journal of Obstetrics and Gynaecology* 111: 133–142.
- Thagard, Paul. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12 (3): 435–502.
- Wason, Peter C. 1968. "Reasoning about a rule." *Quarterly Journal of Experimental Psychology* 20: 273–81.

Journal of Cognition and Neuroethics

Moral Heuristics and Biases

Mark Herman

Bowling Green State University

Biography

Mark H. Herman is a graduate student in philosophy at Bowling Green State University. His primary research interest is the applicability of the heuristics and biases paradigm to moral psychology. His research interests include moral psychology, ethics, rationality, and philosophy of science. He was born and raised in New York City, and received his BA from Wesleyan University. He can be contacted at hermannm_at_bgsu.edu.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Herman, Mark. 2014. "Moral Heuristics and Biases." *Journal of Cognition and Neuroethics* 2 (1): 127–142.

Moral Heuristics and Biases

Mark Herman

Abstract

The cognitive heuristics and biases research program demonstrates the explanatory success of the heuristics-and-biases model. Expanding applications of the model to include moral judgment and action would be in keeping with the model's history of expansion and would further various scientific, philosophical, and practical interests. However, such an application faces a major philosophical challenge: justifying the ascription of "moral erroneousness" within the context of a scientific research program. In other applications of the model, staple bases for ascribing erroneousness include incongruity with empirical facts and violations of ideal theoretical rationality. However, these standards cannot be adapted to the moral domain; this is because as "objective" standards, their moral analogs would involve scientifically inadmissible appeals to "objective" morality. Yet, the previously fruitful use of some "subjective" standards suggests that, at least in principle, a "subjective" moral standard could undergird a fruitful "moral heuristic (-analogues) and biases" research program. While this "subjectivist opening" is promising, it also raises a full slate of further questions for future work.

Keywords

Moral Heuristics, Moral Biases, Heuristics, Biases, Moral Cognition, Metaethics, Moral Rationality, Philosophy, Moral Psychology, Implicit Bias

One death is a tragedy; a million deaths: a statistic.
—Joseph Stalin (attributed)

Most people agree that racial discrimination is morally wrong. Despite this, there are people who discriminate on the basis of race. Why do they do this? To illuminate the project at hand, let us temporarily stipulate that racial discrimination "is morally wrong" (and briefly set aside metaethical worries). With this stipulation, we can ask, "Why do people commit this immoral act?" and ask more generally, "Why do people commit immoral acts?"

Here are some possible explanations. (1) They have the wrong moral view (i.e., they subscribe to a putative morality that permits immoral acts). (2) They do not care (enough) about morality (i.e., the weight they assign to moral considerations is insufficient to preclude decisions to act immorally). And (3) they cannot control themselves (i.e., while they do not sanction their performance of immoral acts, they succumb to impulses—that is, they are weak-willed or *akratic*).

Presumably, there is not a single explanation of all immoral acts. Instead, there is a multitude of explanations, with each explanation applying to its own particular set of cases. Another explanation (or hypothesis) that might merit inclusion amongst this multitude is one that focuses on moral reasoning and appeals to “moral heuristic (-analogues) and biases.”

In the following, I will first, provide a brief primer on heuristics and biases. Second, I will discuss the expanding application of the heuristics-and-biases model and the fittingness of extending that expansion to include moral judgment and action. Third, I will identify potential benefits of that extension and provide some sense of what the application might involve. Fourth, I will introduce a major philosophical challenge confronting this application, namely, justifying the ascription of “moral erroneousness” within the context of a scientific research program. Fifth, I will attempt to make some progress on that challenge and lay out an agenda for future work.

Heuristics & Biases

The notion of “*moral heuristic (-analogues) and biases*” presupposes the conceptual framework of *cognitive* heuristics and biases. Any discussion of the former requires some familiarity with the latter. The following overview, though simplified, should be sufficient.

Humans are constantly making judgments. Some of our judgments are correct; some are erroneous. Some of these errors in judgment conform to patterns. For example, when we are asked to estimate the probability of various events, we tend to *overestimate* the probability of *dramatic* events (e.g., plane crashes) (Tversky and Kahneman 1982a). Errors that fit such patterns are *systematic*; such systematic errors constitute *biases*. The tendency to overestimate the probability of dramatic events is a type of cognitive bias—specifically, a subtype of *availability bias*.

Judgments are the product of cognitive processes. The above judgments regarding the probability of events are the product of a cognitive process that infers the probability of an event *by* assessing how easy it is to recall similar events. In other words, the process uses ease-of-recall as a proxy for probability. Dramatic events, such as plane crashes, are disproportionately easy to recall. As such, assessing ease-of-recall to infer a dramatic event’s probability leads to an overestimate.

We can assess ease-of-recall more efficiently than we can assess probability (as the latter would require executing arduous algorithms). Thus, using the former to infer the latter constitutes a shortcut. Such shortcuts are *heuristics*. Since an event’s ease-of-recall constitutes its *availability*, the process in the above case is called the *availability heuristic*.

Heuristics and biases that involve inferences of this sort are known as “*cognitive*” heuristics and biases.¹

In sum, cognitive heuristics are distinct processes of unemotional inference-making (of the above sort) that constitute cognitive shortcuts (*vis-à-vis* executing ideally rational algorithms). Cognitive heuristics typically yield practical advantages, such as rapid judgment production; however, the judgments they produce are prone to systematic error, that is, *cognitive biases*.

Expanding Application of the Heuristics-and-Biases Model

The principal architects of the cognitive heuristics and biases research program were Daniel Kahneman and Amos Tversky (e.g., Kahneman, Slovic and Tversky 1982; Kahneman and Tversky 1996; 2000; Tversky and Kahneman 2003). The cognitive heuristics and biases model that they employed was adapted from similar models concerning *perceptual* judgments. A particularly influential case was Egon Brunswik’s (1943) application of his lens model to the haze illusion (Kahneman and Frederick 2002, 52; Tversky and Kahneman 2003, 35). The haze illusion is an optical illusion in which objects seen in hazy weather appear farther away than they actually are. This phenomenon was captured as the “perceptual bias” of systematically overestimating the distance of objects in such weather. It was explained by the “perceptual heuristic” of assessing the distance of an object by assessing the perceived clarity of its edges (and the fact that in such weather, perceived edge clarity decreases). This perceptual heritage of Kahneman and Tversky’s model is reflected in the occasional reference to their subject matter as “*cognitive illusions*” (Kahneman and Tversky 1996).

As Kahneman and Tversky’s program progressed, applications of the cognitive heuristics and biases model expanded. This included expanding the range of phenomena captured by the model, and increasing the extent of capture (e.g., advancing from vague awareness of a tendency, to identifying a systematic bias, to hypothesizing a responsible heuristic, to empirically corroborating the heuristic’s existence).

1. The term, “*cognitive*,” may be somewhat of a misnomer. In psychology, “*cognitive*” often regards information-processing, *simpliciter*, and encompasses a variety of capacities, including visual and linguistic capacities (e.g., Anderson 2005; Balota and Marsh, 2004; Eysenck 2001; Kellogg 2002). However, in the context of the “*cognitive*” heuristics and biases program, “*cognitive*” connotes “cold” (i.e., unemotional) “thinking” or “reasoning,” as exemplified by inferences regarding the probability of events. In this sense, the “*cognitive*” is distinct from other kinds of information-processing.

Just as Kahneman and Tversky adapted the “*perceptual* heuristics and biases” model and applied it to probabilistic judgments, Nisbett and Ross (1980) in turn, adapted Kahneman and Tversky’s model and applied it to psychosocial judgments. For example, Nisbett and Ross identified the “act-observer bias,” in which assessments of others’ behavior overestimate the influence of their character (and underestimate the influence of their circumstances). Researchers working in the Kahneman and Tversky tradition went on to (adapt and) apply the heuristics-and-biases model to an expanding range of phenomena, including syllogistic inferences, causal attribution, memory, prediction, and psychophysical judgments. As the presumed bright line dividing reason from emotion was revealed to be more and more blurry, applications of the model expanded beyond solely “cold” (unemotional) reasoning. It incorporated emotion-laden processing, such as the affect heuristic, and reinterpreted “motivated” or “hot” (emotional) irrationality, such as wishful thinking. The expansion of the model’s application to include emotion-induced irrational judgments supports the prospect of further expansion to include *akrasia*-induced irrational judgments and actions. Recent developments in *akrasia* research (e.g., Baumeister and Tierney 2011) appear to render this area increasingly ripe for the identification of biases. Another area of progress is the application of the model to assessments of value (e.g., sunk-cost bias, framing effects, and status-quo bias) (Baron 2008; Connolly, Arkes and Hammond 2000; Gilovich, Griffin and Kahneman 2002; Koehler and Harvey 2004; Schneider and Shanteau 2003).

Importantly, the range of judgments now captured under the heuristics-and-biases model include: emotion-laden judgments, value assessments, and judgments regarding the social domain. These judgments include “ingredients” that in combination, seemingly amount to something fairly close to a moral judgment. Applying a variant of the heuristics-and-biases model to moral judgments (and resultant acts) would be in keeping with the model’s history of increasingly expanding its application.

Moral Heuristics (-Analogues) & Biases

A successful application of the heuristics-and-biases model to the moral domain could further various scientific, philosophical, and practical interests. For one, it would enrich our understanding of human morality—a subject matter of great import and interest. It also would unify disparate, but seemingly related phenomena currently scattered across the literature (such as displaced aggression and, as elaborated below, the identifiable victim effect). It would provide a framework with which similar phenomena could be discovered and/or incorporated into scientific study. And, it could yield implications

for moral and legal philosophy. For instance, it could identify certain immoral acts as attributable to faulty processing (as opposed to for instance, moral indifference); this would have implications regarding moral responsibility.

Perhaps most importantly, a successful program could yield practical benefits. Other applications of the heuristics-and-biases model have yielded *de-biasing* techniques for reducing error. These techniques have targeted biases in: (a) perceptual judgment (e.g., pilots' susceptibility to autokinesis illusions [Civil Aviation Authority 2002]), (b) cognitive judgment (e.g., confirmation bias amongst law enforcement officers [Rossmo 2006]), and (c) psychosocial judgment (e.g., overly-attributing others' actions to character vis-à-vis circumstances [Tetlock 1997]). Applying the heuristics-and-biases model to the moral domain could yield similar preventative and corrective measures to reduce "erroneous" moral judgments, decisions, and actions. These measures could supplement organized efforts and policies to reduce "immoral" actions. Such measures might also have applications within contemporary character education.

Even if the application of the heuristics-and-biases model to the moral domain was unsuccessful, the attempt would still be theoretically constructive. Considering all of the model's fruitful applications, the application of the model itself constitutes a progressive scientific research program (in the Lakatosian sense [Lakatos 2000]). This renders an attempted application to the moral domain a fitting test of the program's boundaries.

Applying the heuristics-and-biases model to the moral domain may reveal the following. (1) Some immoral acts can be explained as instantiations of systematic *moral* errors—that is, "moral biases." (2) Some immoral acts that instantiate moral biases are the product of and thus, explained by "moral heuristics"—that is, distinct psychological processes that constitute moral reasoning shortcuts. Alternatively, (2a) those explanatory processes might be distinguished, not by constituting a shortcut, but by possessing an alternative property that fulfills the same explanatory functions; we could call such processes "moral heuristic-analogues."²

Examples of phenomena that might involve "moral heuristic (-analogues) and/or biases" include: (a) implicit biases (Saul 2013), (b) the influence of displaced aggression upon prejudicial violence (Hovland and Sears 1940), (c) *psychic numbing* (i.e., indifference to the plight of individuals who are 'one of many' in a much greater problem) and its role

2. The explanatory functions of constituting-a-shortcut include (a) endowing heuristics with a telos (and thus, a plausible evolutionary and/or cultural origin) and (b) providing a basis for subsuming similar processes under a scientific kind.

in tolerating genocide (Slovic 2007), and (d) *blaming the victim* (Lerner and Simmons 1966).

Another such phenomenon is the *identifiable victim effect* (Small and Loewenstein 2003). This effect accounts for people's tendency to make greater charitable donations when requests are accompanied by (a) images of particular victims, versus (b) factual statements, even when such statements raise the utility of donating. This difference in donations plausibly constitutes a "moral error;" the systematicity of these decisions would render those "moral errors" instantiations of a "moral bias." A plausible explanation of this "moral bias" is a moral heuristic. Recall that the availability heuristic explained the overestimation of dramatic events by revealing that the estimates were a function not of *likelihood-of-occurring*, but the generally useful, but error-prone proxy, *ease-of-recall*. In this case, a plausible explanation of this difference in donations is that the donations are a function not of *recipients' need* (or something of that sort), but a useful, but error-prone proxy, *emotional impact of the request* (or something of that sort).

It is worth clarifying that the moral erroneousness of interest is error that is "meaningfully *moral*." While this is a slippery notion, some footing can be found by contrasting it with merely logical error committed with moral content. For instance, suppose someone makes a logical error, such as affirming-the-consequent, when responding to a moral dilemma; this would be no more meaningfully a *moral* error, than an irrational gamble on the Kentucky Derby would be an *equine* error. For an error to be meaningfully *moral*, in this sense, it must involve a substantively *moral* mistake. For instance (perhaps), one's attributing moral worth to species of animals on the basis of their cuteness.³

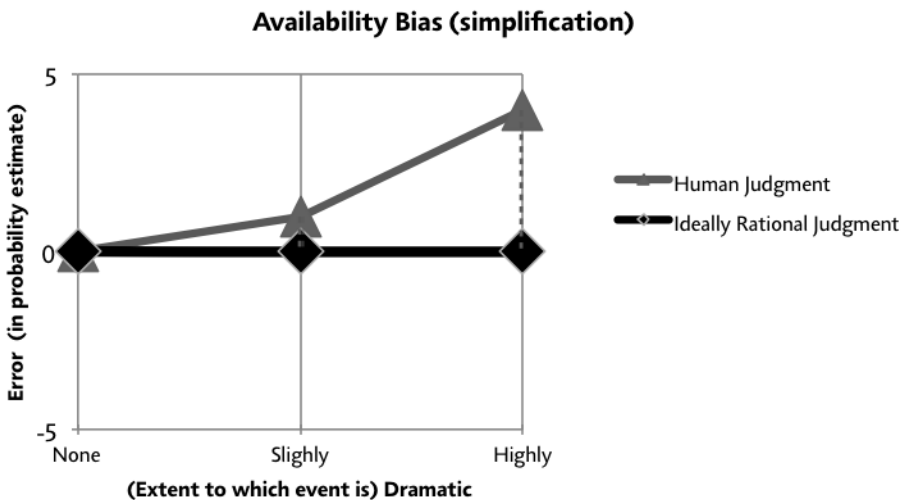
Challenge: Justifying "Moral Erroneousness"

Applying the heuristics-and-biases model to the moral domain faces challenges. One major challenge is: within the context of a *scientific* research program, on what basis

3. Some errors may be attributable to the use of moral content (*à la* Cosmides 1989). However, this is insufficient for being a *meaningfully moral* error. For example, suppose we have a moral reasoning module and it is the only module that affirms-the-consequent. As such, affirming-the-consequent would have an important and interesting association with moral judgment. Nevertheless, such errors would be moral in a different way from those involving *substantively moral* mistakes. It may be that all errors can be interpreted in a way that renders them ultimately reducible to logical errors. Nevertheless, I think we would still be able to make meaningful distinctions between the moral meaningfulness of (a) attributing moral worth on the basis of cuteness, versus (b) affirming-the-consequent with moral content.

can we identify an act or judgment as “morally *erroneous*?” To illuminate this challenge, consider the following.

Identifying a judgment as erroneous implies a standard from which the judgment deviates. With respect to cognitive biases, that standard is ideal rationality. For example, regarding the availability bias, identifying probability estimates as *overestimates* implies a standard of accurate probability estimation; that accuracy is per the dictates of ideal rationality (specifically, probability theory). The essence of the erroneousness ascribed by the availability *bias* is illustrated in the following figure (by the deviation of human judgment from ideally rational judgment).



In contrast with this exemplar, from what standard might a judgment’s deviation warrant the ascription, “*moral error*,” within a *scientific* research program?

To aid the search for an adequate moral standard, it will be useful to consider unproblematic standards from other applications of the heuristics-and-biases model. One unproblematic standard is consistency with empirical facts. Examples of such facts include: the distance of an object (as appealed to by the haze illusion), and subsequent states-of-affairs (as appealed to by various prediction biases, such as the planning fallacy–Buehler, Griffin and Ross 2002).

The unproblematic nature of these error-ascriptions cannot be realized in a scientific research program concerned with moral erroneousness. The above error-ascriptions are for

judgments' inconsistency with facts that ultimately, are inferred due to their providing the most *scientifically epistemically* virtuous account of observations. There is no observation for which a supposed *moral* fact provides the most scientifically epistemically virtuous account to infer *within the context of a scientific* research program.⁴

Another unproblematic standard is deviation from ideal theoretical rationality. For example, error is ascribed to violations of logical principles, such as the conjunction rule (Tversky and Kahneman 1982b). This unproblematic nature cannot be realized with ascriptions of moral erroneousness. This is because within the context of a scientific research program, the case for metaethical uncertainty and skepticism is simply too strong for an unproblematic appeal to an *a priori* morality. In other words, Kant and others' attempts to derive morality from rationality failed (or at least did not succeed to an extent that would license granting moral principles equal standing in science with principles of ideal theoretical rationality) (Kant 1997). For instance, the notion that the scientific admissibility granted to the conjunction rule be bestowed upon the categorical imperative is simply a non-starter.

One might respond that a skeptical case might also be made against ideal theoretical rationality. Perhaps it can, but many of the principles of ideal theoretical rationality that have undergirded error-ascriptions are also indispensable presuppositions of psychological research, if not all of science (e.g., *modus tollens* and the relevance of sample size). As such, rejecting more than a small minority of these principles would amount to a "self-excluding argument." By this I mean that such rejections would exclude one from the community for whom the topic at hand could be of concern (*à la* defending an interpretation of international law by appealing to solipsism). Unlike skepticism of ideal theoretical rationality, moral skepticism leaves science intact (e.g., science is compatible with moral anti-realism, moral nihilism, etc.).

In sum, in the context of a scientific research program, "moral erroneousness" cannot realize the unproblematic nature of the aforementioned error-ascriptions (i.e., per empirical facts or ideal theoretical rationality). Nevertheless, opportunities arise upon realizing that the unproblematic nature of those error-ascriptions is underwritten by their standards' being in an important sense, "objective" or "mind-independent." This sort of error-ascription is not feasible with moral erroneousness because "objective morality" is simply a non-starter within a scientific research program. However, there are unproblematic error-ascriptions in fruitful scientific research programs that rely, not upon

4. This does not necessarily preclude the existence of moral facts; but it does preclude appealing to supposed empirical moral facts within a scientific context.

deviating from “objective” standards, but upon deviating from ideal *practical* rationality or means-to-subjective-ends rationality. Some well-established examples include ascribing error to irrational gambles (Tversky 1969), framing-sensitivity (Baron 2008, 267–71), and scope-neglect (Kahneman and Frederick 2002, 74–75).

For instance, the erroneousness ascribed to an irrational gamble depends upon the presumption that the goal or *end* of such gambling is maximizing expected financial payoff (or something of that sort). However, what grounds privileging this end over ends such as *minimizing* financial payoff, befuddling descriptive rational choice theorists, or having fun? At least within the context of a scientific research program, the privileging of such ends cannot be grounded in the aforementioned sorts of “objective” standards. Such privileging can only be grounded by the gambler’s desire to maximize his financial payoff (or something of that sort). That is, such ends are in an important sense, “subjective.” Upon stipulating such an end, ascribing erroneousness to judgments, decisions, and/or actions that frustrate the realization of that end may very well be based in part, upon deviating from “objective” standards, such as ideal theoretic rationality. For example, the erroneousness of irrational gambles often depends upon principles of probability theory. Nonetheless, the overall standard that grounds the erroneousness of irrational gambles still depends upon the privileging of “subjective” ends and as such, constitutes a “subjective” standard.

One might respond that the “subjectivity” of such standards would render them scientifically inadequate. This is not the case. Roughly speaking, heuristics are distinguished from pertinent contrast class members, such as algorithms, by the property, constituting-a-shortcut. ‘Shortcut’ is an inherently relative concept. In the cognitive program, that which heuristics are shortcuts relative to is: the algorithms dictated by ideal rationality. A fuller expression of heuristics’ distinguishing property is: shortcut-vis-à-vis-ideal-rationality. Ideal rationality does not have a causally efficacious manifestation in the intuitive judgment system; shortcut-vis-à-vis-ideal-rationality is not a causally-relevant property. With respect to the causal workings of the intuitive judgment system, heuristics are indistinguishable from other intuitive processes that happen to conform to ideal rationality (e.g., simple logical rules, such as *ex-elimination*). In short, heuristics are not a natural kind (or some such appropriate analogue). That is, the distinction of heuristics and biases as kinds does not stem from the causal workings of the intuitive judgment system. Instead, it depends upon their relation to ideal rationality, and stems from our valuing ideal rationality as a standard and our choosing to incorporate this standard into a particular classification scheme. While such incorporations merit caution because of certain theoretical and methodological implications, they are ultimately, unproblematic.

They merely reflect the fact that we want to understand things we care about, and that these concerns dictate categories that do not necessarily map onto the strict metaphysical categories reflected in causation. For example, that the category, money, does not map onto such categories does not mean that we should no longer study economics (Fodor 1974). The success of the cognitive heuristics-and-biases program suggests that employing value-driven categories would not impede applying the heuristics-and-biases model to the moral domain.

All in all, the fruitful utilization of “subjective” standards in other applications of the heuristics-and-biases model supports the prospect that “subjective” standards could be similarly fruitful in applications of the model to moral judgment. In other words, we have reason to think that, at least in principle, a subjective moral standard could undergird a fruitful moral heuristic (-analogues) and biases research program.

So, given this “subjectivist opening,” what should that standard be? Great care needs to be taken. Following Kohlberg’s problematic application of Piaget’s model to the moral domain, moral psychology has already witnessed at least one case of a research program mishandling the privileging of particular moral norms (Kohlberg 1958; Piaget 1970; per: Gilligan 1977).

Here are some questions that will need to be answered. Should the standard be based upon cultural relativism (psychology is already replete with it)?⁵ But morality is notoriously controversial; is there sufficient agreement to undergird a particular cultural standard? Even if there is, would such a standard be able to distinguish between moral views that are “erroneousness” from those that are merely atypical (which opposition to slavery once was)? An inability to make such a distinction would make “debiasing” disturbingly Orwellian. Does this make such standards inherently political? Is that problematic?

Relativize to the individual instead? Would moving to the individual level allow for the generalizability that is often essential to the value of research? What would an individualist standard even consist of? Moral principles? Moral competences (*à la* Mikhail’s incorporation of competence vis-à-vis performance—Chomsky 1980; Mikhail 2005)? There is reason to doubt whether individuals’ moral principles (rules or competences) are stable (Zimbardo 2007), genuine (Lichtenstein and Slovic 2006), relevant (Haidt 2001),

5. For example, the discipline’s most authoritative definition of mental disorder (i.e., per the DSM-IV), includes the caveat that behaviors in support of a diagnosis “must not be merely an expectable and *culturally sanctioned* response to a particular event, for example, the death of a loved one” (emphasis added, American Psychiatric Association 2000, xxxi).

accessible (Fischhoff 1991), or even present in any coherent form (Churchland 1996). Should we appeal to the result of some procedure, such as reflective equilibrium (*à la* Rawls 1975)?

Should we instead turn our focus to privileged counterfactual judgments, decisions, and/or actions? If so, which ones? Those of particular deliberative processes (*à la* Singer 2005)? Particular intuitive processes (*à la* Kass 1998)? Those of an idealized self (*à la* Railton 1984)? If so, what kind of idealized self (e.g., Rosati 1996), and under what conditions (e.g., Firth 1952)? Furthermore, how do we operationalize such counterfactual judgments?

In addition, do instances of self-aware immorality (i.e., evil) constitute “moral errors”? Should “moral laziness” be excluded? Should errors-of-omission be excluded? Should the ascription of moral error be limited to explicitly *moral* judgments or include non-moral judgments that deviate from “moral correctness”? And so on.⁶

As one can see, identifying and justifying a standard for ascribing “moral erroneousness” that would be adequate in a scientific research program is a substantial challenge. While the “subjectivist opening” allows for some progress in surmounting this hurdle, there is still much work to be done.

All in all, extending applications of the heuristics-and-biases model to include moral judgments and acts would be in keeping with precedent and offers various benefits. Nevertheless, such an application faces serious philosophical challenges that necessitate future work.

6. The approach I favor is privileging counter-factual judgments. One possible (and preliminary) standard is that judgments are “moral errors” if they meet the following conditions: (1) they are produced and maintained *because of* the agent’s lack of awareness of morally pertinent causes of the judgment, and (2) counterfactually, had the agent had this awareness, it would have prevented or overridden the judgment *because of* the moral beliefs or values of the agent. One benefit of this standard is that preventing these “moral errors” would not require any change in actors’ moral beliefs, values, or motivation.

References

- American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision). Washington D.C.: American Psychiatric Association.
- Baron, Jonathan. 2008. *Thinking and Deciding*, 4th ed. New York: Cambridge University Press.
- Baumeister, Roy F., and John Tierney. 2011. *Willpower*. New York: Penguin.
- Brunswik, Egon. 1943. "Organismic Achievement and Environmental Probability." *Psychological Review* 50: 255–72.
- Buehler, Roger, Dale Griffin, and Michael Ross. 2002. "Inside the Planning Fallacy: The Causes and Consequences of Optimistic Time Predictions." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 250–70. New York: Cambridge University Press.
- Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.
- Churchland, Paul M. 1996. "The Neural Representation of the Social World." In *Mind and Morals*, edited by Larry May, Marilyn Friedman, and Andy Clark, 91–108. Cambridge: MIT Press.
- Civil Aviation Authority. 2002. Fundamental Human Factors Concepts. CAP 719. Safety Regulation Group. <http://www.caa.co.uk/docs/33/CAP719.PDF>
- Connolly, Terry, Hal R. Arkes, and Kenneth R. Hammond, eds. 2000. *Judgment and Decision Making: An Interdisciplinary Reader* (2nd Ed.). New York: Cambridge University Press.
- Cosmides, Linda. 1989. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task." *Cognition* 31: 187–276.
- Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12: 317–45.
- Fischhoff, Baruch. 1991. "Value Elicitation: Is there Anything in There?" *American Psychologist* 46: 835–47.
- Fodor, Jerry A. 1974. "Special Sciences: The Disunity of Science as a Working Hypothesis." *Synthese* 28: 97–115.
- Gigerenzer, Gerd. 1996. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103: 592–96.
- Gigerenzer, Gerd. 1998. "Surrogates for Theories." *Theory & Psychology* 8: 195–204.
- Gilligan, Carol. 1977. "In a Different Voice: Women's Conceptions of the Self and of Morality." *Harvard Educational Review* 47: 481–517.

- Gilovich, Thomas, and Dale Griffin. 2002. "Heuristics and Biases: Then and Now." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 1–17. New York: Cambridge University Press.
- Haidt, Jonathan. 2001. "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108: 814–34.
- Hovland, Carl I., and Robert R. Sears. 1940. "Minor Studies of Aggression: Correlation of Lynchings with Economic Indices." *Journal of Psychology* 9: 301–10.
- Kahneman, Daniel, and Shane Frederick. 2002. "Representativeness Revisited: Attribute Substitution in Intuitive Judgment." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 49–81. New York: Cambridge University Press.
- Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.
- Kahneman, Daniel, and Amos Tversky. 1996. "On the Reality of Cognitive Illusions." *Psychological Review* 103: 582–91.
- Kahneman, Daniel, and Amos Tversky, eds. 2000. *Choices, Values, and Frames*. New York: Cambridge University Press.
- Kant, Immanuel. 1997. *Groundwork of the Metaphysics of Morals*. Mary Gregor, trans. New York: Cambridge University Press. Originally published 1785.
- Kass, Leon R. 1998. "The Wisdom of Repugnance." In *The Ethics of Human Cloning*, edited by Leon Kass and John Q. Wilson. Washington, DC: American Enterprise Institute.
- Koehler, Derek. J., and Nigel Harvey, eds. 2004. *Blackwell Handbook of Judgment and Decision Making*. Malden: Blackwell.
- Kohlberg, Lawrence. 1958. *The Development of Modes of Moral Thinking and Choice in the Years Ten to Sixteen*. Unpublished doctoral dissertation. University of Chicago.
- Lakatos, Imre. 2000. "Falsification and the Methodology of Scientific Research Programs." In *Readings in the Philosophy of Science: From Positivism to Postmodernism*, edited by Thomas Schick, Jr., 20–25. Mountain View: Mayfield.
- Lerner, Melvin, and Carolyn Simmons. 1966. "Observer's Reaction to the 'Innocent Victim': Compassion or Rejection?" *Journal of Personality and Social Psychology* 4: 203–10.
- Lichtenstein, Sarah, and Paul Slovic. 2006. *The Construction of Preference*. New York: Cambridge University Press.
- Mikhail, John. 2005. "Moral Heuristics or Moral Competence?" *Behavioral and Brain Sciences* 28: 557–58.

- Nisbett, Richard E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs: Prentice-Hall.
- Piaget, Jean. 1970. *Genetic Epistemology*. New York: Norton.
- Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13: 134–71.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.
- Rosati, Conni. 1996. "Internalism and the Good for a Person." *Ethics* 106: 297–326.
- Rossmo, D. Kim. 2006. "Strategies to help avoid investigative failures." *FBI Law Enforcement Bulletin* 75.
- Saul, Jenifer. (2013). "Jennifer Saul on Implicit Bias." *Philosophy Bites*, podcast audio, December 7, 2013, <http://philosophybites.com/2013/12/jennifer-saul-on-implicit-bias.html>.
- Schneider, Sandra L., and James Shanteau, eds. 2003. *Emerging Perspectives on Judgment and Decision Research*. New York: Cambridge University Press.
- Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9: 331–52.
- Slovic, Paul. 2007. "'If I look at the mass I will never act': Psychic numbing and genocide." *Judgment and Decision Making* 2: 79–95.
- Small, Deborah A., and George Loewenstein. 2003. "Helping a Victim or Helping the Victim: Altruism and Identifiability." *Journal of Risk and Uncertainty* 26: 5–16.
- Tetlock, Philip E. 1997. "An Alternative Metaphor in the Study of Judgment and Choice: People as Politicians." In *Research on Judgment and Decision Making: Currents, Connections, and Controversies*, edited by William M. Goldstein & Robin M. Hogarth, 657–80. New York: Cambridge University Press.
- Tversky, Amos. 1969. "Intransitivity of Preferences." *Psychological Review* 76: 31–48.
- Tversky, Amos, and Daniel Kahneman. 1982a. "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.
- Tversky, Amos, and Daniel Kahneman. 1982b. "Judgments of and by Representativeness." In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 84–99. New York: Cambridge University Press.
- Tversky, Amos, and Daniel Kahneman. 1982c. "The belief in the 'law of small numbers.'" In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 23–31. New York: Cambridge University Press.

- Tversky, Amos, and Daniel Kahneman. 2003. Judgment under Uncertainty: Heuristics and Biases. In *Judgment and Decision making: An Interdisciplinary Reader*, 2nd ed., edited by Terry Connolly, Hal R. Arkes, and Kenneth R. Hammond, 35–52. New York: Cambridge University Press.
- Zimbardo, Philip. 2007. *The Lucifer Effect: Understanding How Good People Turn Bad*. New York: Random House.

Journal of Cognition and Neuroethics

Reasoning and the Military Decision Making Process

Ibanga B. Ikpe

University of Botswana

Biography

Ibanga Ikpe, Ph.D., teaches Contemporary Analytic Philosophy and Critical Thinking at the University of Botswana and had previously taught at the Universities of Lesotho, West Indies, Jamaica and Uyo, Nigeria. He also served as a Critical Thinking consultant to the Botswana Defence Command and Staff College and still teaches Critical Thinking to students of the college. His research and publications are mainly in the areas of Critical Thinking, Philosophical Analysis and Philosophical Practice. He is an APPA certified philosophical counsellor and a certified conflict mediator.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Ikpe, Ibanga B. 2014. "Reasoning and the Military Decision Making Process." *Journal of Cognition and Neuroethics* 2 (1): 143–160.

Reasoning and the Military Decision Making Process

Ibanga B. Ikpe

Abstract

The archetypal view of the military is that of a hierarchical organization whose members are conditioned to respond to command without question. Its election of obedience as “the supreme military virtue” portrays it as subscribing to the highest degree of group conformity possible within any human organization. This view is not helped by the fact that the military adopts a decision making calculus referred to as the Military Decision Making Process (also referred to as The Estimate Process in some military organizations). This seems to suggest that either the soldiers are not expected to think or that whatever thinking it does passes through a decision-making prism which is devoid of the personal inputs of individual soldiers. The irony of the situation is that a course in critical thinking is a required component of staff school training, which is usually a first step towards command and staff appointments in the military. This paper is an attempt to understand reasoning in the military. Using relevant examples, it examines popular clichés about reasoning and the military and the extent to which they are justified by the structure and function of the military. It also looks at the Military Decision Making Process (MDMP) and the extent to which it supports or inhibits autonomous reasoning by individual soldiers, and compares it with other analytical decision making tools. Drawing examples from specific command and staff decisions it concludes that while regimentation may be appropriate for the rank and file, the capacity for reason is an important asset in the military especially as warfare continues to evolve from the conventional to new and bizarre mutants of war.

Keywords

Military Decision making Process (MDMP), Critical Thinking, autonomous reasoning, military obedience

Introduction

In a scathing caricature of the military, the late African Musician, Fela Anikulapo Kuti referred to them as zombies who will not move, talk or sleep without being commanded to do so but who will do everything, including dying, on command. This perception of soldiers as an unthinking but strictly controlled mob that dispenses violence on command is very common among Africans, especially those who have passed through the painful experience of war or have had direct contact with the military, even in peacetime. It is a perception that started during the European pacification of African tribal groups for colonization and grew through the wars of independence, Africa’s many civil wars and military interventions in politics. In part, this poor reputation may be well-deserved, given the African experience of the brutality of invading colonial armies and the viciousness of

its many militias and national armies. Colonial pacification expeditions were, for instance, notorious for their disregard for the lives of natives, whether such natives were fighting within its ranks or living in the communities that had been slated for pacification. Also, Africa's homemade militias, such as The Lord's Resistance Army (LRA) in Uganda and the Revolutionary United Front (RUF) in Sierra Leone have gained notoriety for their cruelty. On the other hand, national armies, like those of Uganda under Idi Amin, Congo/Zaire under Mobutu and Liberia under Samuel Doe have gone down in history as tools used by unpopular despots to brutalize their citizenry and perpetuate themselves in office. The experience of many Africans has been that soldiering is for the dregs of society; people who cannot fit properly into normal civil life, where reason dictates behaviour. The levity with which military personnel handle human life marks them out as devoid of natural emotions especially so within the African societies where human lives are highly valued and moral value is located, not in the individual, but rather in the *relationship* between individuals. Thus the perception of the military as lacking in reasoning is a product of several years of people's interaction with the military and the tendency of military personnel to operate outside the rules that govern community life and peaceful co-existence.

The perception of the military as an unthinking, destructive but organized violent mob is not restricted to Africa but is also found in academic literature from outside Africa. In discussing military procurement practices for instance, Josip Lučev (2011, 157) observes that "a certain degree of mystification surrounds all military decisions, as if their very existence stands for violence and irrationality incomprehensible to a fully civilized mind, and only justifiable with the harsh realities of the world." For Lučev and perhaps many other people around the world, the military is an unavoidable inconvenience that should be kept away as far as possible from civilized society and only tolerated because its capacity for violence is necessary for keeping the enemies of the state at bay. To this end, the best place for the military is in their barracks or otherwise on a remote and desolate battlefield, far away from other human populations. Such perceptions notwithstanding, one could argue as many have done in the past, that soldiering is in fact a profession of gentlemen; a profession where the virtues of courage, perseverance, endurance and chivalry find their full expression. It could indeed be argued that much of the poor perceptions of the military is due to a profound lack of understanding by the civilian population of the unique nature and function of the military and follows from an attempt to evaluate military behaviour using civilian morality. On the other hand, military theorists like Huntington have argued that the herd mentality that is often associated with soldiers grew out of a failure to distinguish between professional soldiers and enlisted men. Thus for Huntington (1957);

The enlisted men subordinate to the officer corps are part of the organizational bureaucracy, but not of the professional bureaucracy. The enlisted personnel have neither the intellectual skill nor the professional responsibility of the officer. They are specialist in the application of violence not the management of violence. Their vocation is a trade not a profession.

Fotion and Elfstrom (1986, 47) also make this point when they argue, “certainly in the military most draftees and many volunteers can hardly be thought of as professionals. Their superiors might urge them to act in a professional manner or even tell them that they are professionals in the hope that they will act that way.” But it is not evident that the distinction between officers and men makes a difference to people’s perception of the military. This is because the general population does not make that distinction and even if it did, its understanding of the military as a highly hierarchical organization where the behaviour of the rank and file is the responsibility of the officer corps, makes such a distinction immaterial. Again, because it is the enlisted men that constitute the visible part of the military (they are usually the ones who go on patrols, engage the enemy and interact with the population in the course of their normal military assignments), they cannot simply be wished away as Huntington attempts to do. The distinction between professionals and tradesmen, though indicative of the training and hierarchy within the military, cannot be used as an excuse for the violent and often irrational behaviour that people identify with the military. Huntingtons’s restriction of soldiering to the officer corps aims to show that the poor perception of the military by the population is borne out of their interaction with people who should ordinarily not represent the military. But Michael Thiesfeld, in his rebuttal of Huntington makes a case for the inclusion of enlisted man within the cadre of professionals. He makes this point when he observes that “young Soldiers are patrolling on foot, engaging the populations, making life and death decisions in a matter of seconds, and are often conducting these activities without a Commissioned Officer looking over their shoulder” (Thiesfeld 2010). Thus, Thiesfeld argues that professionalism should not only be attributed to the officer corps but should also be extended to the enlisted men. This notwithstanding, it is important to note that the enlisted men are always under the command of officers, who, as professionals, ought to direct them to behave appropriately. Thus, irrational behaviour, even by enlisted men cannot be excused under the pretext that they are not professionals, rather, it should reflect on the officer corps and the military in general.

The behaviour of officers and men is not the only reason for the perception of many that the military is an organization where the autonomous reasoning is suppressed. The hierarchical structure of the military, within which individuals appear to have little opinion concerning their profession, actions or existence, reinforces this view. Greg Foster (2004, 91), for instance, observes that “the military is, by nature, a hierarchically organized, authoritarian institution built on rank, the sanctity of command, uniformity and rigid rule following.” Sam Sarkesian (1981, 12) makes the same point when he observes that “personal value systems, institutional requirements and community perspectives will never be in perfect harmony in terms of military professionalism” (Cockerham, and Cohen, 1980,1273). The above gives the impression that individual reasoning and the independent consideration of facts are not encouraged, especially when such reasoning deviates from what has become conventional. More often than not, an officer or enlisted man is required to suspend his critical judgment in the choice of alternatives and rather “use the formal criterion of the receipt of a command or signal as his basis of choice.” This portrays the military a regimented organization that discourages critical judgment and personal initiatives. Again the fact that obedience is often cited as the “supreme military virtue” (Huntington 1957, 74) supports the view that the autonomous reasoning is not a feature of the military. This is because unquestioning obedience leads to regimentation; a situation where agents do not necessarily understand why certain orders have to be obeyed nor are they encouraged to reason concerning such orders.

Critical Thinking and the Military

The above view of the military notwithstanding, the military often prides itself as a rational organization where critical thinking is encouraged and officers are expected to rationally assess command and staff challenges and respond to them with reasoned solutions. They are always eager to show that a professional soldier is not merely someone who has acquired a level of proficiency at military manoeuvres and the use of military equipment, but rather, one who possesses “certain skills and perhaps even a sense of responsibility to exercise these skills in certain ways and at certain times” (Fotion and Elfstrom 1986, 48). Thus, even when a commander is tasked with a particular military objective, the expectation within the military is that he still has the responsibility to reason as to how best the objective could be achieved within the confines of relevant laws and best international military practices. This view of soldiers as strategic thinkers is underscored by Cardon and Leonard as follows:

In an era of persistent conflict, our Army requires versatile leaders, critical and creative thinkers capable of recognizing and managing the myriad transitions necessary to achieve success. In a dynamic and complex situation, these include not just friendly transitions but those of adversaries as well as the operational environment. Commanders and staff must possess the versatility to operate anywhere - along the spectrum of conflict and the vision to anticipate and adapt to transitions that will occur over the course of an operation. (Cardon and Leonard 2010, 4)

In other words, Cardon and Leonard are of the view the contemporary military environment is ever changing and as a result throws up challenges which the professional soldier has to confront creatively and rationally. But it is not only the techniques and materials of warfare that are always changing; the rules of engagement, temperament of the contemporary soldier and the operational environment also change such that the old requirement that a soldier should merely "obey the last order" no longer holds. Experience at trouble spots around the world has shown that the work of the soldier is not restricted to fighting and inflicting losses on the enemy but includes other functions which may require non-military skills and the appreciation of dictates of other cultures. In the current military environment, soldiers are required to think on the move and in the process make decisions that reflect the interests of their country, the safety of their men and the dictates of international conventions. It is in recognition of this need that most military colleges seek to improve the reasoning capacity of their student-officers by electing a course in Critical Thinking as a core component of staff college education. The need for a course in critical thinking arises out of the fact that, as Emilo (2000, vi) observes, "the current educational system has not prepared us for tomorrow's challenges. We've been taught what to think but not how to think." This is why there is a belief in many military training establishments is that military training should not follow this trend and instruct military professionals on what to think but rather should help them develop critical thinking. Mead (2013, 12) confirms this when he observes, "For the first time in its history, the military wants to teach even junior personnel not just *what* to think but *how* to think." Thus, in making Critical Thinking a part of the staff college curriculum, the military hopes to "provide conditions favourable for the development of the autonomous personality" (Szasz 1970, 142) and through this help to develop the capacity for autonomous reasoning within the officer corps. What is not clear, however, is whether in referring reasoning they mean the same thing as Walton (1990, 401) when he

says, “in a critical discussion, logical reasoning can be used where one party, in dialogue with another party, tries to convince this other party that his (the first party’s) point of view is right.” Again, it is not clear whether in referring to autonomy they share the views of Kant that it “is not merely self-assertion or independence, but rather thinking or acting on principles that defer to no ungrounded authority” (O’Neill 1992, 289–299).

Although a course in critical thinking is a required component of staff college education, it is doubtful whether student-officers actually get to develop their critical thinking capabilities in the way that is envisaged in the curriculum. It is also doubtful whether the development of such capacities takes place in an atmosphere where the autonomous reasoning is allowed to flourish. The reason for this doubt is that staff college training takes place within a rigid and time-critical environment where student officers are required to work at various educational tasks that combine formal tactical military training with normal academic work. In such a crowded milieu, what passes as the development of critical thinking is sometimes merely an instruction on the tools of critical thinking. It is not evident that an attempt is ever made to ensure that such tools are applied, or that there is a link between formal academic instruction in critical thinking and the practical business of warfare. Again although the military prides itself as encouraging the application of reason towards the achievement of military objectives, it puts in place step-by-step decision making procedure to which officers are expected to adhere in making military decisions. This procedure, referred to as the Military Decision Making Process (MDMP) has been described as “an indispensable model for the problems posed by a bipolar security environment” (Cardon and Leonard 2010, 2). Thus we have an ironical situation where on the one hand the military seeks to promote autonomous reasoning and on the other, seeks to control the type of reasoning that gets done by setting out the parameters for such reasoning. Rather than encourage reasoning, dictating the way reason gets done could actually be said to inhibit reasoning. This view that the MDMP inhibits reasoning is not readily accepted by the military; rather they look upon it as an instrument that ensures the application of reasoning in military decision making. Their understanding is that, setting the parameters within which reasoning gets done ensures that officers do not make rash decisions that are based on the whims of the moment but would rather be forced to go through a process that guarantees some reasoning.

The Military Decision Making Process

The Military Decision-Making Process is used by many military establishments as a standardized reasoning calculus to ensure precision and uniformity in military decision

making. It is an analytical tool “employing a time-intensive, but logical sequence to analyze the situation, develop a range of options, compare these options, and then select the option that best solves the problem” (Marr 2001, 8–9). It involves the effort of the commander and his staff-officers in bringing to bear their collective cognitive resources toward achieving a military objective, with the men and material available to the commander at the time. It emphasizes the importance and expertise of staff-officers and the opinion they bring to the discussions that produce the blueprint of the mission. Such staff officers as are in intelligence, logistics, air support, artillery, infantry and others with specializations relevant to the mission all contribute to the decision from the perspective of their expertise and the resources they control. In employing this calculus there is a belief that if sustained and appropriate reasoning is applied to a military objective, such an objective could be achieved efficiently. The MDMP consists of seven steps, viz:

- Step 1 - Receipt of Mission
- Step 2 - Mission Analysis
- Step 3 - Course of Action Development
- Step 4 - Course of Action Analysis
- Step 5 - Course of Action Comparison
- Step 6 - Course of Action Approval
- Step 7 - Orders Production

Each of these steps consists of various tasks, the completion of which constitutes the full application of the MDMP. Apart from Steps 1 and 7 which involve the mere communication of information, the remaining steps of the process ought to involve purposeful and reflective judgment which is the hallmark of critical thinking. In developing the MDMP the military considered the enemy to be a thinking, innovative and unpredictable adversary who will employ every guile in pursuing his objective. It is considered that the only way to overcome a resourceful enemy is to get into his thinking curve, understand the drivers of his thought and take steps to frustrate his plans. In doing so, critical thinking is usually regarded as an indispensable tool and its role in military strategy is underscored by Cardon and Leonard when they argue:

Critical thinking also helps distil the immense amounts of information and determine those elements of information that are most relevant to the situation. This is an important step in mitigating the risk associated with guidance that does not fully account for the complexities of the operational environment. Critical thinking helps to clarify guidance and

enables commanders to achieve a mutual understanding of the current situation and the desired end state. (Cardon, and Leonard 2010, 6)

It could therefore be argued that in principle, the MDMP encourages the use of Critical Thinking in the hope that it will ensure a thorough analysis of the enemy and the combat environment. But what is accepted in principle may not always translate into practice. Thus, it is important to determine whether or not in practice, the MDMP promotes reasoning.

The MDMP, just like other analytical decision making tools, conceives of decision making as a series of analytical steps which when properly followed lead to appropriate decisions. Its two basic components consist of identifying/understanding the problem and implementing the solution is also common to other analytical decision making processes. For instance, the Critical Thinking Decision Making Process (CTDMP), proposed by Anne Thomson (1999, 92–3) has options, information, consequences and evaluation as its essential components. The development of options which is the first component of Anne Thomson's CTDMP, comes in as task 2 of the third step of the MDMP and could be said to belong to the initial stage of problem identification/understanding. The requirement by Thomson's CTDMP that we seek information as a second step in decision making comes in as step 4 of the MDMP. The consideration of the consequences which constitutes step 3 of the CTDMP features as task 1 of step 5 in the MDMP, while 'evaluation' which ends the CTDMP, apart from recurring consistently at the different steps of the process constitutes task 2 of step 5. Whereas the first and second components of the CTDMP could readily be classified as identifying/ understanding the problem, the last stage of the process could be associated with implementing the solution. From the above, one could argue that the MDMP is as good as any analytic decision making tool and may even be better since it incorporates features that are not found in other such instruments.

The above notwithstanding, it is important to note that the MDMP is a time consuming decision making instrument which, ironically is meant to be applied in a time-critical military environment. This presents decision making challenges to the commander because, often, the time that is available for decision making, is not usually adequate for a full application of the MDMP; this is especially so when the decisions involved are time-critical field decisions. A full application of the MDMP requires that the commander and his staff perform 41 tasks between the receipt of the mission and the issuing of Warning Order (WngO) for the mission. Each of the tasks is of a technical nature and may require interaction with other units, friendly forces and enemy forces. The tasks also have to be accomplished within the timeline set by higher headquarters

and delimited by the commander. It is not surprising, therefore, that many commanders and staff have complained about the time and resources that go into using the MDMP. Marr for instance, observes that “unit performance at the U.S. Army’s combat training centres (CTCs) suggests that tactical units have difficulties in applying the MDMP” (Marr 2001, 2) and part of the reason for this is because they are too long for use in real combat environments. This view is corroborated by Garcia (1993, 3) when he claims that “observations from subject matter experts observing staffs during training indicate that they have difficulty conducting the military decision-making process.” Although Garcia does not say why staff officers have these difficulties, there is no doubt that much of it has to do with the conflict between the time for initiating an action and the time it takes for the process to be completed. Thus, even where a commander earnestly wishes to adhere to the MDMP, common sense will dictate to him that such a decision will arrive too late to ensure the success of an operation. Thus it is not uncommon for the commander to switch from the MDMP, which is an analytical decision making tool, to an intuitive decision making process which is not so time intensive. This defeats the whole purpose of the MDMP which has always been to prevent officers from acting intuitively rather than rationally.

Again, although the MDMP is usually presented as a reasoning calculus, some of the tasks required by the process are of a practical nature, such that, subjecting them to sustained reasoning would be superfluous. Upon receiving a WngO from higher headquarters at the first step, for instance, the MDMP tasks for the commander are mainly routine. He is expected to;

- Alert the staff.
- Gather the tools: Higher Head Quarters order, Maps, Standard Operating Procedure (SOPs), Appropriate Manuals, and Running estimates
- Update running estimates.
- Conduct initial assessment

It is apparent that these processes do not require sustained reasoning and to apply such reasoning to them would make a mockery of the process. The second step in the process, mission analysis, consists of 17 tasks, viz:

- Analyse higher Headquarters order.
- Perform initial Intelligence Preparation of the Battlefield (IPB).
- Determine specified, implied, and essential tasks.
- Review available assets.
- Determine constraints.
- Identify critical facts and assumptions.

- Perform risk assessment.
- Determine Commander's Critical Information Requirements/Essential Elements of Friendly Information (CCIR / EEFI).
- Determine initial Intelligence, Surveillance, and Reconnaissance (ISR) plan.
- Update operational timeline.
- Write the restated mission.
- Deliver a mission analysis briefing.
- Approve restated mission.
- Develop initial Commander's intent.
- Issue Commander's planning guidance.
- Issue warning order.
- Review facts / assumptions.

Although the above include tasks that appear to entail sustained reasoning, it is important to decide whether, like Benard Gert, we believe that "rationality is not purely procedural, that is, there is no specifiable procedure of deliberation that can plausibly be said to confer rationality on whatever goals might happen to emerge from it" (Gert 1991, 103). If we do, then the fact that the MDMP is a structured and specified procedure for reasoning condemns it. Also, the fact that officers are limited to a specific operational timeline that is either set by the commander or by Higher Headquarters makes it more likely that they will run through the different steps rather than give them reasoned consideration. Again it is important to note that a Commander may, at his discretion, decide whether to do the full MDMP or to abbreviate the process after receiving the WngO. In abbreviating the process, the commander may wish to adopt any four of the techniques detailed in force manual 101-5. These include:

- a. Increase the commander's involvement, allowing him to make decisions during the process without waiting for detailed briefings after each step.
- b. The commander to become more directive in his guidance, limiting options.
- c. The commander to limit the number of COAs developed and war-gamed.
- d. maximize parallel planning.

In each of these abbreviated options, it is the tasks and steps that involves sustained reasoning and analysis that get jettisoned.

Although the MDMP is often cited as an analytical reasoning instrument for the military, the realities of military life appears to stand in the way of its effective application. The realities referred to here, begin early in military training where a concerted effort is made to replace the individualism of civilian life with a military groupthink and a socialization process is put in place to suppress the autonomous personality and

replace it with a heteronomous one. Early in their training, cadet officers and men are constantly reminded of the virtues of obedience and made to appreciate the need to defer to the superior knowledge and experience of the commander, trusting that the net effect of carrying out his/her command will be beneficial to all concerned. A culture of obedience is important, not only because the military must be united in confronting an objective but also because such a unity of purpose translates to efficiency and efficacy. This culture of obedience is sometimes carried over when officers are given command and staff appointments and is sometimes seen as a disincentive for critical judgments and independent opinions. Some officers would rather recycle a judgment made by their superiors in similar circumstances or adopt a position from the military operational manual than make a critical judgment of their own. Thus instead of making a reasoned and unique judgment that effectively addresses the particular situation, the MDMP allows officers to pretend that they are actually making a reasoned judgment whereas they are merely going through the motions.

Another disincentive for critical judgment in early career officers is the need to avoid blame for operational failures. A failure to achieve an objective sometimes spells catastrophic outcomes for the formation and can weigh heavily on the officer responsible. Blame for such failure would be mediated if the decision was based on ideas that emanate from the rule book or from what has been done in the past but would be severe if it was a novel idea that emanates from critical judgment of the officer. In such cases, the officer would be adjudged to be lacking in judgment and incapable of making sound decisions. This is to say that critical judgment and the resultant new approach to a military objective is fine, so long as it achieves results and since no one can say for certain when a critical judgment will achieve such results, officers are more likely to make "safe" decisions and that is, decisions that are based on the training and indoctrination of the particular military. Thus although, in making decisions, commanders and staff are expected to be guided by professional judgment gained from experience, knowledge, education, intelligence and intuition, many officers (especially those that are new to command responsibilities) sometimes shy away from taking this step and instead try to second guess the kind of decision that their superiors would expect in the circumstance, or stick to what has worked best in the past. Here again, the MDMP aids and abets such recycling of old decisions by putting in place tasks that could be performed and labelled as reasoning without necessarily applying critical judgment and innovation.

An argument for saying that the MDMP is a disincentive to reasoning comes from the fact that most officers do not look at the MDMP as an invitation to reason concerning the task at hand but as reason itself. Their relationship to the MDMP is best understood

using the distinction made by A. H. Simon between programmed and non-programmed decision making. According to Simon (1977, 46), "Decisions are programmed to the extent that they are repetitive and routine, to the extent that a definite procedure has been worked out for handling them so that they don't have to be treated from scratch each time they occur." Programmed decisions are usually approached from the standpoint of organizational policy and the rules for, and specific ways of handling them are usually well known within the organization. In the military for instance, officers and men are fully aware of the standard procedure for 'intelligence preparation of the battlefield (IPB),' viz., 'determining specified, implied, and essential tasks,' 'reviewing available assets,' 'determining constraints,' 'identifying critical facts and assumptions,' and so on. An attempt to introduce something novel or extraneous to such an established operating procedure would appear uncalled for and potentially destabilizing to the general routine of the military. Since such decisions are programmed, they are not subjected to as much discussion or consideration as they ordinarily should, rather they are automated to ensure consistency and also save time. On the other hand, decisions are non-programmed "to the extent that they are novel, unstructured and unusually consequential." Such decisions do not typically follow established guidelines and their rules are complex and little understood. In the military, such decisions are required once in a very long while and are often the exclusive preserve of very senior officers who may or may not consult their subordinates in the course of making the decision. Given its structure and function, it is clear that the MDMP is a programmed decision making tool and merely functions as a checklist which officers run through when making decisions. The lofty ideals that are the hallmark of critical thinking are usually lost to the automated check listing that is the hallmark of the military.

Of Reasoning and Warfare

The popular clichés concerning reasoning and the military notwithstanding, there is no doubt that the difference between a successful army and a mediocre one lies in the quality of the command decisions taken by its officers. Although the unschooled may believe that soldiering consists merely in aiming a gun at a target and shooting, the fact that military engagements entail strategizing should be ample evidence that soldiering is an art which require the full application of the faculties of reasoning. Indeed, "military strategy consists of the establishment of military objectives, the formulation of military strategic concepts to accomplish the objectives and the use of military resources to implement the concepts" (Lykke Jr. 1977, 186). Any military strategist will confirm

that a good strategy begins with having sufficient respect for the enemy, understanding him as a rational and calculating being and approaching confrontation with him with a careful analysis of his motives, capacities, emotions and history. They understand that it is very rarely that chance plays a role in military success. Most especially they understand that high quality military decisions have never been a product of mere chance but rather of sustained reasoning, especially when military decisions deviate from the norm. It is the quality of command decisions that gives a numerically inferior army advantage over a numerically superior one, as was the case in the Falkland war. The three prong strategy of declaring a Total Exclusion Zone (TEZ), deploying small but highly mobile raiding parties and landing troops at San Carlos bay, 50 miles from Port Stanley, was a carefully thought out decision. Although the techniques involved are not new, the decision to employ them within the particular context of the Falklands involved a careful juxtaposition of ideas and options.

It is also the quality of the command decisions that could give a rag tag army advantage over a technically proficient army. An example of this was the decision of the Iraqi Republican guard to avoid direct confrontation with a technically superior US led military coalition but rather fight a psychologically debilitating war of attrition with them from within the population; turning what should have been a clinically precise military operation into attrition warfare. Here again, although the concept of guerrilla warfare is not entirely new, the decision of the Iraqis to adopt it at that time and especially their election to deploy Improvised Explosive Devices (IED) as an important component of the overall plan was truly ingenious. Again, the Vietnamese general Vo Nguyen Giap is often referred to as military genius because of his effective military decisions, many of which are still shrouded in mystery. His signature tactic of attacking several enemy interests at the same time did not only force the French to disperse their numerically and tactically superior force into smaller units that the Vietnamese forces could easily engage but was also psychologically debilitating for the Americans when, many years later, they sort to defend south Vietnam from communist rule. Vo's other strategy of timing every offensive to achieve the greatest negative public opinion impact in the enemy country became his specific contribution to the techniques of war. Williams (1963, 130) for instance notes that "the assault on Dien Bien Phu, was clearly timed to coincide with the 1954 Conference at Geneva where the Indochinese territories were to be partitioned and parcelled out." News of the brutality of the attack not only had an impact on French public opinion concerning the war but also on the public opinion of other countries with overseas colonies. Its impact was such that, it is suspected that President Eisenhower's ignored the recommendations of his war cabinet on sending troops to help the French

fight communism in Vietnam because of the negative public opinion that the attack had in the United States.

This two prong strategy was also used with devastating effect in the Tet offensive which, despite not achieving its set objective, is considered a turning point of American involvement in the Vietnam War. According to Guan (1998, 346), “contemporary American intelligence reports which have shaped much of the writings on the Tet Offensive laid emphasis on Giap’s opposition to Thanh’s strategy for a quick and decisive victory, preferring the continuation of the protracted war strategy.” It is generally believed that Giap’s preference for a protracted battle was due to the expected impact of such a battle on the morale of American soldiers on the field and on American public perception of the war. Attacking several cities at the same time gave the impression that the Americans were fighting an enemy with an endless supply of men while the siege on American forces gave the impression that the Americans were easy targets for the enemy. It is the propaganda value of his attacks that is often cited as the reason for American withdrawal from Vietnam such that, “Osama bin Laden and other terrorists have routinely mentioned Vietnam as a model for the type of victory they are seeking, a debilitating blow to the American will that results in demoralization at home and withdrawal of troops abroad” (Robbins 2010, 52). This is despite the fact that Vo’s strategy involved a colossal loss of men and the victory was achieved at a great cost to the nation’s productive capacities. Decisions such as these are not the product of regular military processes (indeed Giap was not a professionally trained soldier), rather it is the product of sustained application of reason to the task of defeating the enemy.

Conclusion

Facing an enemy in a situation where the fortunes of the men in uniform and indeed an entire country depends on the decision of a few is very eerie and should be approached with all the intellectual resources that are available to the officers and men. Reducing such intellectual resources to a checklist of items such as on the MDMP can never stand any army in good stead as it confronts an enemy. This, especially so, since the elite military colleges around the world open their doors to cadets and student officers from many nations, thus ensuring that ‘the run of the mill’ operational procedures that are taught at military colleges are no secret but can be easily recognized by the enemy who may use this knowledge to his advantage. Again the fact that nations change alliances ever so often results in a situation where officers and men who, until recently, were fighting from within the same trench may be planning on how to annihilate one another. What this

means is that the military training and other assistance that a country gives to a friendly nation at some point may become a lethal weapon in the hand of an enemy when such a nation turns round to be hostile. Since programmed decision making tools such as the MDMP form part of such assistance, it is important for officers to develop the capacity for non-programmed decision making sure that they have an advantage when confronted with such an enemy. This is not to say that the MDMP should be completely abandoned but it is saying that it should not be relied upon as an “indispensible tool.” Also, the socialization of officer cadets should not completely strip them of their individuality and thereby their capacity to bring innovative ideas into the conduct of war. Officers and men should be encouraged to develop and maintain their capacity for rational enquiry and by extension their ability to be innovative and creative, even in the application of the MDMP. Limiting the capacity of officers and men to effectively use their reason in the application of the MDMP put the military at a disadvantage when they face a resourceful enemy and this may escalate the human and material cost of winning a war. Developing a capacity for rational thought thus becomes indispensable especially in an ethical military that can ill afford the waste its human and material resources.

References

- Cardon, Brigadier General Edward and Lieutenant Colonel Steve Leonard. 2010. "Unleashing Design: Planning and the Art of Battle Command." *Military Review* 92 (4): 2–12.
- Cockerham, William and Cohen, Lawrence. 1980. "Obedience to Orders: Issues of Morality and Legality in Combat among U.S. Army Paratroopers." *Social Forces* 58 (4): 1272–1288.
- Emilio, George A. Major. 2000. "Promoting Critical Thinking in Professional Military Education." Research Report Submitted to the Faculty, Air Command and Staff College, Air University, Maxwell Air Force Base, Alabama, U.S.A.
- Foster, Greg. 2004. "Obedience as a Failed Military Ethos." *Defense & Security Analysis* 20 (1): 90–96.
- Fotion N., and Elfstrom, G. 1986. *Military Ethics, Guidelines for Peace and War*. London: Routledge and Kegan Paul.
- Garcia, Maj, Jacob A. 1993. "The Requirement for an Abbreviated Military Decision-Making Process in Doctrine." Master Of Military Art And Science Diss., Fort Leavenworth Military Academy.
- Guan, Ang Cheng. 1998. "Decision-Making Leading to the Tet Offensive (1968) - The Vietnamese Communist Perspective." *Journal of Contemporary History* 33 (3): 341–353.
- Huntington, Samuel P. 1957. *The Soldier and the State: The Theory and Politics of Civil-Military Relations*. New York: Vintage Books.
- Lučev, Josip. 2011. "Convergence in Military Procurement Practice: Responses to Asymmetry." *Politička misao*, Vol. 48 (5): pp. 157–172.
- Lykke, Arthur F. 1997. "Defining Military Strategy." *Military Review* 77 (1): 182–186.
- Marr, Major John. 2001. *The Military Decision Making Process: Making Better Decisions Versus Making Decisions Better*, unpublished monograph, School of Advanced Military Studies, United States Army Command and General Staff College Fort Leavenworth, Kansas, U.S.A.
- O'Neill, O. 1992. "Vindicating Reason." In *The Cambridge companion to Kant*, edited by P. Guyer, 280–308. Cambridge: Cambridge University Press.
- Postow, B. C. 1991. "Gert's Definition of Irrationality." *Ethics* 102 (1): 103–109
- Robbins, James. 2010. "An Old, Old Story." *World Affairs* 173 (3): 49–58.
- Sarkesian, Sam. 1981. *Beyond the Battlefield: The new Military Professionalism*. New York: Pergamon Press.

- Simon, H.A. 1977. *The New Science of Management Decision*. Englewood Cliffs, NJ: Prentice-Hall.
- Szasz, Thomas A. 1970. *Ideology and Insanity: Essays in the Psychiatric Dehumanization of Man*, Garden City, NY: Anchor Books.
- Thiesfeld, Michael Maj. 2010. "Who is a Professional." Paper presented to the 2010 Fort Leavenworth Ethics Symposium. Available online at http://c.ymcdn.com/sites/www.leavenworthethicssymposium.org/resource/resmgr/2010_General_Papers/Thiesfeld.pdf.
- Thomson, Anne. 1999. *Critical Reasoning in Ethics*. London: Routledge.
- Walton, Douglas N. 1990. "What is Reasoning? What Is an Argument?" *The Journal of Philosophy* 87 (8): 399–419.
- Williams Lea E. 1963. "Review of George K. Tanham, *The Military Doctrines of Mao Tse Tung Applied in Vietnam Communist Revolutionary Warfare, the Vietminh in Indochina*." *Journal of Southeast Asian History* 4 (2): 128–133.

Journal of Cognition and Neuroethics

The Role of Emotional Intuitions in Moral Judgments and Decisions

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, 2014. Volume 2, Issue 1.

Citation

Gee, Catherine. 2014. "The Role of Emotional Intuitions in Moral Judgments and Decisions." *Journal of Cognition and Neuroethics* 2 (1): 161–171.

The Role of Emotional Intuitions in Moral Judgments and Decisions

Catherine Gee

Abstract

Joshua D. Greene asserts in his 2007 article “The Secret Joke of Kant’s Soul” that consequentialism is the superior moral theory compared to deontology due to its judgments arising from “cognitive” processes alone without (or very little) input from emotive processes. However, I disagree with Greene’s position and instead argue it is the combination of rational *and* emotive cognitive processes that are the key to forming a moral judgment. Studies on patients who suffered damage to their ventromedial prefrontal cortex will be discussed as they are real-life examples of individuals who, due to brain damage, make moral judgments based predominately on “cognitive” processes. These examples will demonstrate that the results of isolated “cognitive” mental processing are hardly what Greene envisioned. Instead of superior processing and judgments, these individuals show significant impairment. As such, Greene’s account ought to be dismissed for does not stand up to philosophical scrutiny or the psychological literature on this topic.

Keywords

Consequentialism, deontology, moral judgments, emotion, cognition, ventromedial prefrontal cortex

Recent research in cognitive science has shed new light on philosophical moral theories by unveiling what cognitive processes are at work when we contemplate moral issues. Joshua D. Greene asserts that consequentialism is the superior moral theory due to its judgments arising from “cognitive” processes alone without (or very little) input from emotive processes. However, I disagree with Greene’s position and instead argue it is the combination of rational and emotive cognitive processes that are the key to forming a moral judgment. Furthermore, I will show that even with possible amendments, Greene’s account does not stand up to philosophical scrutiny or the psychological literature on this topic.

I will begin by briefly detailing Greene’s argument and contrasting it with the claim that both emotion and reason are part of the moral judgment cognitive process and are very much intertwined. In the second section, I will demonstrate what happens when one only uses “cognitive” processes when forming moral judgments as Greene endorses, by presenting research on patients with ventromedial prefrontal cortex (VMPFC) damage. As a result of brain damage, these patients appear to primarily utilize the “rational” or “cognitive” regions of the brain with little input from emotive signaling. This is similar

to how Greene asserts one arrives at a consequentialist moral decision. However, studies have shown that decisions made via these isolated “cognitive” reasoning processes can have disastrous results when implemented into action. Finally, I will conclude with a discussion of the material in which I will be arguing why a more complete and accurate moral theory will recognize the role that emotional processes have in our moral decision making, and why an ideal theory cannot omit an emotive aspect as Greene posits.

1. Greene on Consequentialism and Deontology:

Greene argues that deontology emphasizes moral rules, often in terms of rights and duties, and one makes this type of judgment by following these moral rules. While deontology does take consequences into account, it may require us to do things that do not produce the best possible consequences (Greene 2007, 37). However, for consequentialism, Greene asserts “the moral value of an action is in one way or another a function of its consequences alone” (2007, 37). It aims to produce the best overall consequences, if not directly then indirectly as consequences are the *only* things that ultimately matter (2007, 37). Greene states that deontology and consequentialism refer to *psychological natural kinds* and instead of philosophical inventions the two terms are better classified as “philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking” (2007, 37). What he is interested in here are the “relevant functional roles” of the two theories and not the “conventional philosophical definitions” (Greene 2007, 38). These functional roles evoke different kinds of judgments, which then lead to different types of conclusions based on these judgments. Consequentialist judgments favor “characteristically consequentialist” conclusions, such as it is better to save more lives when faced with a moral dilemma like the trolley problem¹ (Greene 2007, 39). Deontological judgments, in contrast, are judgments that are in favor of “characteristically deontological” conclusions like “it’s wrong despite the benefits” (Greene 2007, 39). Greene admits that while he would not assert that either approach is strictly emotional or “cognitive”, he argues that consequentialism is more cognitively driven while deontology is more emotionally motivated (2007, 41).

1. There are several variations of this problem but the basic idea is this: there is a runaway trolley racing down a track headed towards five people who are unable to move. You happen to be standing next to a lever that if pulled, will divert the trolley to a side track, thus saving the five people. However, you notice that there is one person on the other track who is also unable to move. The problem is deciding what you ought to do in this situation: Pull the lever and save the five people but one person will be killed as result of the diverted trolley, or do nothing and save the one but the trolley will kill the five.

Greene and his colleagues used the trolley problem to test the cognitive versus emotive distinction and to discover in which types of scenarios participants utilized cognitive or emotive mental processes. They proposed that the footbridge dilemma² in the trolley problem is more of an “up-close” and “personal” matter to deliberate that “is more emotionally salient than the thought of bringing about similar consequences in a more impersonal way (e.g., by hitting a switch)” (Greene 2007, 43). When harm is impersonal, argues Greene, it fails to trigger alarm-like emotional responses and thus allows one to respond in a more “cognitive” manner that is more detached than personal cases (2007, 43). However, when one contemplates personal moral dilemmas (the more “up-close” scenarios) Greene argues there is relatively greater activity in the emotion-related areas of the brain,³ whereas contemplation of impersonal moral dilemmas resulted in relatively greater neural activity in the “cognitive” areas of the brain.⁴

Greene states that deontological judgments tend to arise from emotional responses and as such deontological philosophy is to a large extent an exercise in moral *rationalization* rather than being grounded in moral *reasoning* (2007, 36). He contrasts this with consequentialism, which he argues arises from quite different psychological processes, processes which are more “cognitive” than those that lead to deontological judgments, and as such are “more likely to involve genuine moral reasoning” (Greene 2007, 36). Greene states “cognitive” representations are inherently neutral because they “do not automatically trigger particular behavioral responses or dispositions” (2007, 40). In contrast are the “emotional” representations that have automatic effects and are thus behaviourally valenced (Greene 2007, 40). Greene uses “cognitive” to signify the term being used in a narrow sense that is in contrast with emotion (2007, 40).

Greene takes his argument a step further and posits that our emotional intuitions have been “shaped by morally irrelevant factors having to do with the constraints and circumstances of our evolutionary history” and thus thinks they ought not to be trusted (2007, 75). Instead, the better bet is to uphold consequentialist principles that are devoid of these morally irrelevant intuitions as they “provide the best available standard for

2. Similar to the original problem, however this time you are standing on a footbridge above the runaway trolley headed towards the five people. There is a fat man standing next to you on the bridge and you realize that he is big enough to stop the trolley if he fell onto the tracks in front of it. Do you push him to stop the runaway trolley, thus killing him but saving the five, or do you do nothing and the trolley will continue on and kill the five?

3. The posterior cingulate cortex, the medial prefrontal cortex, and the amygdala (2007, 44).

4. The dorsolateral prefrontal cortex and inferior parietal lobe (2007, 44).

public decision making” (Greene 2007, 77). I will return to this argument later in the paper to explore it further after I present additional research for discussion.

In contrast with Greene is Jonathan Haidt, a proponent of the Social Intuitionist Model, which argues that when one encounters a moral dilemma one first feels an intuition, such as a flash of disgust or discomfort, and “knows intuitively that something is wrong” (Haidt 2001, 814). It is not until there is a social demand for justification for the intuition that one attempts to create an argument to defend their initial gut reaction or intuition. Haidt asserts that “one becomes a lawyer trying to build a case rather than a judge searching for the truth” (2001, 814). According to the Social Intuitionist Model, a moral judgment is “caused by quick moral intuitions” and is followed by slow moral reasoning after the fact (Haidt 2001, 817). Haidt argues that reason alone is not enough to motivate one to act morally and asserts that “[r]eason can let us infer that a particular action will lead to the death of many innocent people, but unless we care about those people, unless we have some *sentiment* that values human life, reason alone cannot advise against taking the action” (2001, 816). The contrast this model makes between intuition and reasoning is not the contrast between emotion and cognition that Greene takes to be the case. Instead, for the Social Intuitionist Model, “[i]ntuition, reasoning, and the appraisals contained in emotions are all forms of cognition” (Haidt 2001, 818). Intuition occurs quickly, effortlessly, and automatically whereas reason occurs more slowly and requires some effort (Haidt 2001, 818). Moral judgments are the result of automatic and effortless moral intuitions and moral reasoning is the outcome of an effortful process that begins after the moral judgment is already made and arguments are required to support this existing judgment (Haidt 2001, 818).

2. Studies on Ventromedial Prefrontal Cortex Damage

Neurobiological research has been conducted in an attempt to understand the mechanisms of intuitions and the findings are consistent with the Social Intuitionist Model’s marrying of emotive and rational processes. The ventromedial prefrontal cortex (VMPFC) is the area located behind the bridge of the nose (Haidt 2001, 824) that triggers somatic states from one’s memories, knowledge, and cognition (Bechara and Damasio 2005, as cited in Sobhani and Bechara 2011, 643). Damage to this area results in very similar behaviour to that of a psychopath. Damasio’s somatic marker hypothesis “states that experiences in the world normally trigger emotional experiences that involve bodily changes and feelings” (Haidt 2001, 825). Once the brain is “properly tuned up” by repeated experiences of these emotional conditionings (think Pavlov) “the brain areas that monitor

these bodily changes begin to respond whenever a similar situation arises” (Haidt 2001, 825). These emotional signals “function as covert, or overt, biases for guiding decisions” (Damasio 1994, as cited in Sobhani and Bechara 2001, 643). Where the VMPFC⁵ comes into play is that it integrates these feelings, or ‘somatic markers’, with the individual’s other planning and knowledge functions and helps the brain decide quickly on a response that is the result of both emotional and rational processes (Haidt 2001, 825). Damage to the VMPFC leads to defective activation of somatic states (the emotional signals) which in turn affects the perceived value of given scenarios and options (Damasio 1994, as cited in Sobhani and Bechara 2011, 643).

The link between the functioning of the ventromedial area of the prefrontal cortex and moral behaviour has been explored extensively by Damasio and his colleagues. Patients with brain damage that is restricted to the VMPFC show no reduction in their reasoning abilities (Damasio 1994, as cited in Haidt 2001, 824). In fact, “[t]hey retain full knowledge of moral rules and social conventions, and they show normal abilities to solve logic problems ... and even hypothetical moral dilemmas” (Damasio 1994, as cited in Haidt 2001, 824). However, when the patients are presented with real decisions “they perform disastrously, showing poor judgment, indecisiveness, and what appears to be irrational behavior” (Haidt 2001, 824). According to this research “the central deficit resulting from destruction of the VMPFC is the loss of emotional responsiveness to the world in general and to one’s behavioral choices in particular” (Haidt 2001, 824). For example, when shown pictures that elicit a strong reaction from control subjects without VMPFC damage (such as nudity, mutilation, or death) the patients with the brain damage report feeling nothing in response to the pictures and skin conductance responses confirm their lack of emotional response (Damasio, Tranel, and Damasio 1990, as cited in Haidt 2001, 824).

Researchers also conducted studies involving gambling tasks and the results demonstrate that patients with VMPFC damage are lacking the unconscious biases that are derived from previous experiences with reward and punishment (Sobhani and Bechara 2011, 645). These biases help deter people from pursuing a course of action that is not advantageous in the future as they learn from their past mistakes (Sobhani and Bechara

5. While the VMPFC is a critical component, other regions in the neural system ought to be noted as well, such as the amygdala, insula and somatosensory cortices, dorsolateral prefrontal cortex, and hippocampus (Sobhani and Bechara 2011, 642). Furthermore, different brain regions may also provide different contributions to the overall process of decision-making (Bechara and Damasio 2005, as cited in Sobhani and Bechara 2011, 642). The involvement of these other regions do not, however, undermine the importance the role the VMPFC plays in moral decision making.

2011, 645). Without these biases one still may possess knowledge of what is right and wrong (or the best course of action in the gambling task) but without the biases the knowledge alone is “not sufficient to ensure an advantageous behavior” (Sobhani and Bechara 2011, 645). The researchers conclude that having knowledge in absence of the emotional or somatic signaling leads one to experience dissociation between what one knows and how one decides to act (Sobhani and Bechara 2011, 645). Sobhani and Bechara hypothesize that without the emotional component in the moral judgment decision-making process, a person is “left making a more pragmatic decision based on the facts of the situation, with a special emphasis on the outcome of the situation and less so on the inferred or abstract events or intentions” (2011, 647). For example, patients with VMPFC damage judged *attempted harms*, including attempted murder, as more morally permissible relative to the controls on a scenario task used to judge harmful intent (Sobhani and Bechara 2011, 647). The patients demonstrated that they did not factor in the negative intentions described in the scenarios and instead focused on the action’s neutral outcome.

This is essentially the consequentialism described by Greene, but instead of commending it as a preferable “cognitive” approach Sobhani and Bechara view it as an impairment resulting from a lack of emotional input via the VMPFC. Thus, despite the difference in opinion regarding the outcome both sets of researchers can be seen to be arguing for the same conclusion, that a lack of emotion leads to a consequentialist position. To reiterate, Greene defines consequentialism as the superior moral theory over deontology due to the former resulting in judgments and decisions that are unencumbered by one’s emotions. This permits one to make a moral decision that is the result of “genuine moral reasoning” rather than a rationalization based on emotional intuitions, as in deontology. A real life example of this type of decision making occurs in individuals with VMPFC damage who make “consequentialist” decisions, both moral and otherwise, due to a lack of input from somatic signaling.

It is interesting to note the differences that arise in cases of ventromedial prefrontal cortex damage that occurred later in life compared to injuries that were sustained at a very young age. Those with early-onset lesions “failed to show normal learning of rules and strategies from repeated experience and feedback” and had significant impairment of social-moral reasoning and verbal generation of responses to social situations (Anderson et al. 1999, 1033). These patients “demonstrated little consideration of the social and emotional implications of decisions, failed to identify the primary issues involved in social dilemmas and generated few response options for interpersonal conflicts” (Anderson et al. 1999, 1033). Their performance was “in stark contrast” to patients who had adult-onset

prefrontal lesions, as the latter were still able to access the “facts of social knowledge” in laboratory scenarios (Anderson et al. 1999, 1033).

3. Discussion

Hopefully, by this point, the problems that arise from Greene’s preference of “cognitive” processing that is isolated from emotional input are apparent. By examining studies on ventromedial prefrontal cortex damage we are able to see real-life examples of Greene’s “cognitive” processing in action, and the result of decision making when emotion is divorced from reason. Decision-making can still occur, as we have seen in these cases, but it is devoid of the valuable emotive and somatic marker inputs that lead to truly “moral” judgments. The somatic markers are what motivate and guide us towards a moral decision based on the judgment they evoke. If a purely “rational” process, that is a process absent the somatic markers, was what led us to make the “best” or most rational moral judgments as Greene argues, then patients with ventromedial prefrontal cortex damage would arguably be the best example of this theory in practice. Unencumbered by emotional processing that struggles with how personal and “up-close” a moral dilemma is, these individuals should be making the clearest, most concise, and correct moral decisions. However this is not the case, as these individuals do not learn from repeated mistakes even with explicit knowledge of the consequences of their decisions (Sobhani and Bechara 2011, 642). They demonstrate impaired judgment and decision-making as well as impaired moral judgment (Sobhani and Bechara 2011, 642). Individuals who suffered the brain damage in their adulthood began to make choices that were no longer advantageous to themselves and these decisions were “remarkably different from the kinds of choices they were known to make before their brain damage” (Sobhani and Bechara 2011, 642). They often decided against their own best interests and many suffered a loss of family and friends as well as their social standing (Sobhani and Bechara 2011, 642). These dramatic results were not due to a decline in intelligence, for in “striking” contrast to their real-life decision making impairment, these patients performed normally in most laboratory tests of problem solving, and conventional clinical neuropsychological tests demonstrated the patients’ intellects were still in the normal range after the damage (Sobhani and Bechara 2011, 642).

These research results stand in sharp contrast to Greene’s claims that I discussed in the first section of this paper. If intuitions are morally irrelevant and the emotive drive tied to them ought to be disregarded, then why do patients with ventromedial prefrontal cortex damage not make advantageous decisions instead of disastrous ones? After all, these

individuals are real life examples of agents who are able to think and act the way Greene endorses, yet the studies show how impaired their moral judgments and decisions are. I do not think the problem lies with consequentialism itself but rather Greene's treatment and definition of the theory. While consequentialism looks to the consequences of an action to determine if it is morally permissible or not, nothing necessitates that the examination of the consequences ought to be done by omitting emotion entirely. When one looks at an attempted murder case, for example, Greene's treatment of it appears to be similar to the viewpoint of the patient with VMPFC damage in that the intention does not matter for in the end, a life was not actually lost. Most would disagree with this conclusion as it feels like something is missing. *Feels* is the key word here, as one experiences intuitive discomfort at excusing the attempted act just because no concrete physical consequence resulted. Haidt makes my point rather well when he says "[i]t is not contrary to reason to kill your parents for money unless it is also contrary to sentiment" (2001, 824). By removing the emotive aspect from consequentialism Greene is omitting an important cognitive input that is necessary for a better and more complete account of consequentialism.

As mentioned in the beginning of the paper in the section on Greene's argument, he throws in an attempt to either weaken his claim or hand-wave about the divisibility of cognition and emotion by asserting that he does not believe that either approach is strictly emotive or "cognitive." He continued in the same sentence that he also does not maintain that "there is a sharp distinction" between the two (2007, 41). Greene states that he is sympathetic to Hume's claim that moral judgments (including consequentialism) must have an emotional component to them (2007, 41). However, he follows this statement by saying "[b]ut I suspect that the kind of emotion that is essential to consequentialism is fundamentally different from the kind that is essential to deontology, the former functioning more like a currency and the latter functioning more like an alarm" (Greene 2007, 41). Can Greene make such assertions and still maintain his position? I argue that he cannot. The separation he finds in cognition and emotion play too important of a role in his argument for it to be eliminated or minimized, for his conclusion requires this distinction. Greene's favouring of consequentialism over deontology is based on the mental process he finds superior—those of "cognitive" origin. This process cannot be superior to emotive processes due to the cognitive's "genuine moral reasoning" if there is any element of the emotive processes involved. The way Greene has presented his argument is that any involvement of automatically triggered behavioural or emotive responses would seem to taint the colder "rational" cognitive process by allowing "rationalization" to occur based

on “morally irrelevant factors.” Thus, Greene’s argument only stands if there is a strict separation of the cognitive and emotive processes.

However, the research shows—and Greene even admits—that there is no such divorce of the two mental processes in individuals with ‘normal’ brain function. Furthermore, as I have demonstrated, when this separation does exist in brain-damaged patients the moral judgments and conclusions are hardly what Greene envisioned. As a result, Greene is left in a difficult position. He could amend his account to fully deny the possibility that moral judgments include both cognitive and emotive processes, in which case his argument itself will be stronger due to his conclusion necessitating this distinction. However, by taking this firmer stand Greene will be in opposition to the numerous studies in the psychological literature that disprove the isolation of the two processes. On the other hand, Greene could amend his account to accommodate the research and accept the intertwined roles cognitive and emotive processes play in moral judgments. In doing so, however, he cannot maintain his conclusion that consequentialism (as he defines it) is the superior moral judgment due to it arising from the “better” cognitive process. None of the available options (leave his account as is, strongly maintain the separation, or accept the marriage of the two types of processes) are viable, and unless Greene can come up with another amendment that I have overlooked, his account ought to be dismissed.

Moral judgments cannot be made without *both* rational and emotive processes or it no longer remains a moral judgment. It would then become something else, the result of a different type of judgment and decision that stemmed from impaired mental processing. Emotional intuitions that arise from somatic markers are the key which distinguishes mere mental processing and decision making from *moral* cognitive processing and decision making. Any attempt to separate cognitive from emotional processing contradicts the psychological literature on the matter as well as the essence of what it means to judge morally.

References

- Anderson, Steven W, Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio R. Damasio. 1999. "Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex." *Nature Neuroscience* 2 (11): 1032–37.
- Greene, Joshua D. 2007. "The Secret Joke of Kant's Soul." In *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development*, edited by Walter Sinnott-Armstrong, 35–79. Cambridge: MIT Press.
- Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34.
- Sobhani, Mona and Antoine Bechara. 2011. "A Somatic Marker Perspective of Immoral and Corrupt Behavior." *Social Neuroscience* 6 (5-6): 640–52.

Journal of Cognition and Neuroethics

Asking for Reasons as a Weapon: Epistemic Justification and the Loss of Knowledge

Ian Werkheiser

Michigan State University

Biography

Ian Werkheiser is a doctoral student in the Department of Philosophy at Michigan State University, with specializations in both Environmental Philosophy & Ethics and Animal Studies. His research areas include environmental philosophy, social and political philosophy, and epistemology (particularly social epistemology). His dissertation will focus on the capabilities approach and food sovereignty. It will argue that community epistemic capacity is a necessary requirement of meaningful political participation, particularly in issues around food and environmental justice. Ian is currently involved in several collaborative projects. He works with Dr. Paul Thompson on the Sustainable Michigan Endowed Project, a foundation dedicated to increasing research into sustainability in the state of Michigan. Over the summer he was a co-PI on an NSF-supported Long-Term Ecological Research project, "Recognizing Value Pluralism among Ecosystem Services Experts and Public Stakeholders." This year he is co-organizing the second annual Workshop on Food Justice at MSU, a workshop which brings together academics and activists.

Acknowledgements

The author wishes to acknowledge Dr. Kristi Dotson for her help with earlier drafts of this paper.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Werkheiser, Ian. 2014. "Asking for Reasons as a Weapon: Epistemic Justification and the Loss of Knowledge." *Journal of Cognition and Neuroethics* 2 (1): 173–190.

Asking for Reasons as a Weapon: Epistemic Justification and the Loss of Knowledge

Ian Werkheiser

Abstract

In this paper, I will look at what role being able to provide justification plays in several prominent conceptions of epistemology, and argue that taking the ability to provide reasons as necessary for knowledge leads to a biasing toward false negatives. However, I will also argue that asking for reasons is a common practice among the general public, and one that is endorsed by “folk epistemology.” I will then discuss the fact that this asking for reasons is done neither constantly nor arbitrarily, but rather in a systematic way that produces ignorance by oppressing some knowledge and some knowers, in particular those from already marginalized groups. After looking at the implications of all this, I will ultimately argue that we must be very careful when we ask for reasons, and acknowledge it as the powerful weapon it is.

Keywords

Ignorance, Epistemic Violence, Epistemic Justice, Epistemic Silencing, Justification, Reason-giving, Contextualism, Social Epistemology, Externalism

I. Introduction

Asking people “How do you know that?” is a common response to hearing new information, and in fact, might be seen as a good epistemic habit before accepting what someone has to say. In this paper, however, I argue that this question is more problematic than it seems. Since, on the externalist view, having good reasons for belief is not a requirement for knowledge, and since even on many internalist views having good reasons *at the moment* is not a requirement for knowledge, demanding these reasons can illegitimately lead to a failure of knowledge uptake, and in extreme cases can get the original speaker to abandon knowledge she would otherwise have had. Given that there is reason to believe that this question is systematically asked more often of ideas that challenge the *status quo* and asked of people who come from marginalized groups, this asking for reasons is not only epistemically problematic, but is an example of epistemic injustice (Fricker 2007) and epistemic violence (Dotson 2011). In this way, this paper can be seen as contributing to the growing discourse on the production and maintenance of

ignorance (e.g., Tuana & Sullivan 2006; Alcoff 2007), in particular the “invested ignorance” that maintains unjust social relations (Townley 2011).

In this paper, I will look at what role being able to provide justification plays in several prominent conceptions of epistemology, and argue it is a form of over-determination of knowledge, but one which is often not recognized as such in ordinary discourse. I will then show how taking this over-determination as necessary for knowledge leads to a biasing toward false negatives, and that this is done neither constantly nor arbitrarily, but rather in a systematic way that oppresses some knowledge and some knowers. After looking at the implications of this, I will ultimately argue that we must be very careful when we use the weapon of asking for reasons.

II. Justification

There are many definitions of justification in epistemology, so let me begin by establishing the kind of justification I will be using here. For the purposes of this paper, I do not need too rigorous a definition of justification, but it cannot be merely any X that is added to true belief to make it knowledge. Rather, I am talking about the kinds of justification that might be called “good reasons” for our beliefs, such as Laurence Bonjour’s (1978) conception of justification in his (1978) article. He says there that “Cognitive doings are epistemically justified, on this conception, only if and to the extent that they are aimed at this goal [truth] – which means roughly that one accepts all and only beliefs which one has good reason to think are true” (113).¹ For those who subscribe to something like access internalism in their epistemology, meaning that we must have cognitive awareness of these reasons, it is clear what role justification (in this sense) plays: without good reasons for our beliefs, we would not have knowledge. I will use “justification” and “reasons” interchangeably in this paper.

In some strict versions of internalism, such as Bonjour’s, it is necessary to be able to produce these reasons on demand: “A person for whom a belief is inferentially justified need not have explicitly rehearsed the justificatory argument in question to others or even to himself. It is enough that the inference be available to him if the belief is called into question by others or by himself . . . and that the availability of the inference be, in the final analysis, his reason for holding the belief” (110). Other versions of internalism, which we might call lax internalism, do not require that one always be able to provide reasons at any moment in which one has knowledge. Instead they require only that, on

1. All page numbers for this article are taken from its occurrence in Sosa et al. (2008) *Epistemology: an Anthology*.

reflection (perhaps quite involved reflection), a subject will come to know her reasons for her belief, and thus be able to provide them. As Robert Audi says in his (2001) article, “The idea is largely that a person with a justified belief *has* a justification for it, in the sense of grounds one can adduce in giving a justification in an appropriate context, such as one in which a puzzled friend asks why one believes the proposition in question. *Giving* a justification (in the relevant sense) requires having the justifier(s) in consciousness, but being *able* to give it by citing the relevant grounds does not. . . . Grounds are a kind of resource: we can genuinely possess resources without having them in our hands at the time” (23).

If one is an externalist, however, what this kind of justification is “for” is not as clear. In externalism, reasons are not necessary for knowledge as long as people are using a reliable process to form belief, or there is a law-like relationship between their beliefs and the truth, or some other externalist conception of warrant. Nevertheless, it is the case that people often can and do provide reasons for how and why they know something (whether or not these are their actual reasons; something we will discuss in section IV), and asking for these reasons is something that non-philosophers at least sometimes do. In the next section, we will look at two representative examples of externalist accounts that offer an opinion on internally having reasons.

Ila. Reasons as Epistemically Uninteresting Phenomena

The first approach is to acknowledge that giving reasons is a thing that humans do, perhaps even something that is constitutive of our second nature as socialized animals (e.g., McDowell 1995). However, though it is a common human practice, in this account it is not one which is epistemically interesting, or one that epistemologists ought to concern themselves with. A good example of this is John Greco’s account. As Greco says in his (2005) on the purpose of epistemic evaluation,

We care about whether Mary is, in general, a responsible and reliable cognitive agent. We also care about whether, in this instance, Mary arrived at her belief in a reliable and responsible way. We also care, of course, about whether Mary’s belief is true. These are important considerations about Mary and about her belief – considerations that are important from the point of view of information-using, information-sharing beings such as ourselves. (267)

For Greco, these are all externalist questions of warrant, and not internalist questions of justification, because the subject will often not have access to these epistemic considerations.

As for justification, on the other hand, he asks “Why should we care that Mary is no more blameworthy for her belief at the moment than she was the moment before?” (267). Further, “Why would we be interested to know whether [a belief] *b* is licensed by norms of evidence that *S* accepts? Why would this be an important evaluation to make?” (268, emphasis in the original). Greco is not claiming that these are not things that people try to determine in a conversation, but only that from an epistemic point of view (one interested in truth conducivity) internal justification is not interesting (260).

IIb. Reasons Being Favorable to Truth

A relatively recent treatment of justification in a framework that also contains external considerations is Alston’s, in which he rejects the term “justification” in favor of “epistemic desiderata.” Nevertheless, his third category of desiderata maps on fairly well to what we have been calling “justification:”

6. *S* has some high-grade cognitive access to the evidence, and so on, for [a belief] *B*.
7. *S* has higher-level knowledge, or well-grounded belief, that *B* has a certain positive epistemic status and/or that such-and-such is responsible for that.
8. *S* can carry out a successful defense of the probability of truth for *B*.
(Alston 2005, 43)

For Alston, these desiderata are not directly truth-conducive, but unlike for Greco they are useful (or “interesting”) epistemically, because:

They contribute to *S*’s being in a position to arrange things in a way that is favorable to acquiring truth rather than false beliefs. . . . The basic point is that the more we know, or are able to know, about the epistemic status of various beliefs or kinds of beliefs, the better position we are in to encourage true beliefs and discourage false beliefs. . . . [These desiderata promote] the ability to distinguish between beliefs that are likely to be true and those that are not and to encourage the development of those that satisfy the former description. (44)

Thus, what we have been calling justification is on Alston's account neither necessary nor sufficient, but is a nice over-determiner for a true belief, in that it makes that belief more likely to be true and lets us be surer of it. For Alston, justification is also nice for someone to have often, because it encourages our ability to develop true beliefs.

Thus, in internalism, justification is always good and will ideally always be present, though on the lax account it does not always need to be immediately accessible, while in the dominant treatments of justification within externalism,² justification can be seen as either helpful or superfluous. In none of the accounts we have looked at has providing or asking for reasons been seen as dangerous for knowledge. I will argue in the next section, however, that if justification is taken to be important or even necessary for knowledge, then it loses this benign character.

III. Asking For Reasons

As we saw in the previous section, justification may have a role to play in an external account of knowledge, or at the least can just be seen as a non-epistemic human activity, but it is neither necessary nor sufficient for knowledge. In internalist accounts, on the other hand, it is the *sine qua non* for turning true beliefs into knowledge, though not all accounts require it to be immediately available in all instances. When we look at how requiring justification plays out over the set of a person's total true beliefs, then, we see that for any account other than strict internalism, there will inevitably be fewer things that we can provide reasons for at the moment than the set of what we know. This means at least that we will miss instances of knowledge if we only count those beliefs that we can currently justify.

IIIa. False Negatives

Consider the following cases:

1. Some of our beliefs were made by reliable processes, but we no longer remember the processes, and so cannot provide good reasons to ourselves or others for these beliefs. An example of this might be someone who learned how to speak a language in school, and so used to know the explicit rules for conjugating into

2. Perhaps not surprisingly, the account that is more sympathetic to justification, Alston's, is billed as a compromise between externalism and internalism. If Greco is right and internalism believes that all aspects of knowledge other than truth are internal, and externalism believes that at least some elements are not, Alston's account would definitely be an externalist one. Whether it is an externalist account or an account with external elements is not relevant to our discussion here, however.

a particular tense, but now simply knows the correct conjugation for a particular word, but when they try to teach the language to others they realize that they cannot recall the rule that generates the conjugation.

2. Some of our beliefs are based on very reliable processes (such as perception), but we have no conscious access to the cues we are using and so again cannot provide good reasons. An example of this phenomenon would be facial recognition, a task we are very adept at performing, but one which we often cannot describe. Most of us at least are at a loss to explain how we recognize our mother from among similar-looking people.
3. Some of our beliefs are based on very reliable processes, but we do not have conscious access to what these processes are or how they work and so cannot say if they are reliable. We are consciously aware of the output of these processes, but cannot give reasons as to how we came to have them. An example of this might be the intuitions that come from experience, for example believing (correctly) that someone is being dishonest without being able to explain why.
4. Finally, some of our beliefs might be non-linguistic, and formed through (reliable) non-linguistic processes. As such we are not able to linguistically give any reasons for the belief, or even to adequately explain the belief. A good example of this might be a baseball player who knows where to position the bat in order to hit a ball, and can even show you how to do it (given enough time for embodied practice), but cannot explain exactly how she knows it, or what exactly she is doing.

In most externalist accounts, all of these would be counted as knowledge since they are based on reliable processes, and giving justification is not required. What about internalist accounts? Here the question is a bit more murky. 1 would definitely count as knowledge for what I have been calling lax internalism, if the speaker were able over time to reconstruct the rule. 2 counts as knowledge in many versions of internalism, because though we cannot explain all of the features we use to make visual judgments (or at least not without extensive research into our cognition), we can at least say that we are using perception and that the image appears “motherly” to us. 3 and 4 may well not count as knowledge for many versions of internalism, depending on how much awareness a given version required, and how extensively the person is allowed to study themselves and still have it count as “internal” access. Nevertheless, it is at least possible for some accounts of internalism that at least some of these examples count as knowledge. Leaving aside professional epistemology, it seems intuitive that at least some of the above examples or analogues of them would usually be called knowledge.

If one demands that the subject provide reasons, however, then presumably none of the above would count as knowledge. This would depend on how deep the requirement for reasons went, because the person in each case could give some sort of first-level explanation, such as “I figured it out a long time ago” in 1, or “I think he’s lying based on my years of experience talking to people” for 3. Nevertheless, if a questioner persisted in demanding more complete reasons from a semi-skeptical position, all four examples would fail fairly quickly.

Thus, asking for reasons as a test for knowledge biases us toward false negatives—many things that are in actual fact knowledge are not counted as such. This is particularly true if one subscribes to an externalist account, but even lax internalists such as Audi who do not require people always being able to immediately provide reasons will lose some things they would like to call knowledge.

If neither version of epistemology absolutely requires providing reasons, at least on demand, why worry about it? The answer is that this test of asking for reasons is very common in our society. Philosophical epistemology (other than strict internalism) may not require being able to give reasons for one’s beliefs in a given moment, but what we might call “folk epistemology”—the way we unreflectively think about knowledge in our culture—does. Asking people for reasons is seen as a very good idea in situations with high stakes, or when the person’s claim is very surprising, or when the person does not seem entirely trustworthy.

Or rather, we *sometimes* act as if we needed reasons to have knowledge. On the one hand, we often say that we know things which we cannot justify, as in the examples 1 through 4 above. On the other hand, asking how we know something, or why we think something, is a common occurrence in normal conversation, at least in some situations. Thus at times we seem to endorse a need for being able to give reasons, while, at other times, we clearly find it unnecessary. It is also clear that one of these is viewed by most people as a higher standard than the other, namely the limited class of justifiable beliefs. While we often say that we know something without being able to give reasons for it, we usually do not explicitly state that we have no reasons; to do so feels like admitting that we do not “really” know it at all, but merely believe it. Once the conversation brings up the need for reasons, our standard for knowledge changes. When this fact is combined with the “false negatives” idea above, the result is that *we can lose knowledge by having to provide our reasons for it*.

If this begins to sound like contextualism, that is because contextualists have given a good model for how standards of knowledge can change in different contexts. In the next section we will look at two prominent models of contextualism, and we will see

that while they do a good job of describing the dual intuitions we have about needing reasons, they overestimate our ability to switch back from a context which requires giving reasons to one that does not. As a result, they must be modified to deal with the force doubt can have.

IIIb. Contextualism

IIIb1. DeRose's Account

A good example of contextualism can be found in DeRose's (1995), in which he presents a contextualist solution to skepticism. He argues that

When it is asserted that some subject *S* knows (or does not know) some proposition *p*, the standards of knowledge (the standards for how good an epistemic position one must be in to count as knowing) tend to be raised, if need be, to such a level as to require *S*'s belief in that particular *p* to be sensitive for it to count as knowledge. (36)

When skeptics bring up their challenges, the standards of knowledge are raised infinitely high because the skeptical hypotheses are chosen so as to never be sensitive. Thus when we entertain these hypotheses, nothing can count as knowledge (36). This is seen as a useful account because it explains the intuitive appeal of the skeptical challenge, without threatening the truth of our ordinary claims, which are usually stated in contexts of lower standards for knowledge, for, as DeRose says, "the fact that the skeptic can install very high standards that we don't live up to has no tendency to show that we don't satisfy the more relaxed standards that are in place in more ordinary circumstances and debates" (37).

IIIb2. Lewis's Account

Like DeRose, David Lewis also attempts to address skepticism by presenting a contextualist rule for when we can ascribe knowledge, in his (1996) article. His definition of knowledge ascription is that "Subject *S* knows proposition *p* iff *p* holds in every possibility left uneliminated by *S*'s evidence; equivalently, iff *S*'s evidence eliminates every possibility in which not-" (551). He later adds a clause *sotto voce* which he contends is always included in idiomatic (non-logical) uses of the word "every": "*S* knows that *p* iff *S*'s evidence eliminates every possibility in which not-*p*—Psst!—except for those possibilities that we are properly ignoring" (554). His response to the skeptical challenge lies in the way he cashes out "properly ignoring."

After giving many rules for when we are “properly” allowed to ignore possibilities, he includes a rule for when we *cannot* ignore a possibility, which he calls the “*Rule of Attention*” (559). In this rule, whatever possibilities are in our mind cannot be properly ignored (559). Lewis calls this “More a triviality than a rule” (559), but actually this “triviality” is his response to skepticism. This rule shows that when a skeptic presents scenarios like being a brain in a vat (which we would otherwise have ignored following one or more of his previous six rules), we can no longer ignore them, and so do not know things we did know before these possibilities were raised to our attention (559–560). As with DeRose’s account, Lewis’s version of contextualism allows him to explain the force of skeptical hypotheses without giving everything away to skepticism. We can still say that we know many everyday things, just not in a conversation that includes the possibility of the skeptical scenarios.

IIIb3. Contextualism and Asking for Reasons

Both of these accounts can be applied to the topic of this essay, though we will need to modify them slightly. When people ask for reasons, they are not precisely asserting that there is some p that the subject does or does not know, nor that there are possibilities that one was not previously considering in which the p is false. Nevertheless, they are still upping the stakes and changing what the requirements are for knowledge in the context of the conversation.

Recall that we know many things that we cannot (at least at the moment) explain with good reasons, but we also (in our folk epistemology) agree that we should be able to provide justification for any knowledge we actually have. Yet, even that strict version of folk epistemology does not require that we always *entertain* those reasons; that would be an impossibly high standard. So as long as our attention is not drawn to the question of whether we can provide good reasons for our knowledge, we still hold it. Once the question is asked, however, we evaluate our knowledge, and if we find that we cannot provide these reasons, then we no longer know it.

IIIb4. Switching Contexts

This kind of loss of knowledge is only a temporary problem in most versions of contextualism, but in actual conditions, changing back from a context of high epistemic standards to one with lower epistemic standards is not so easy.

For DeRose, “As soon as we find ourselves in more ordinary conversational contexts, it will not only be true for us to claim to know these very O ’s [typical knowledge claims such as “I have hands”] that the skeptic now denies we know, but it will also be wrong for

us to deny that we know these things" (5). Thus, we can move back and forth between standards of knowledge easily depending on our conversational context. For Lewis, it is slightly harder to change our context, because it is a possibility we are considering in our mind rather than just a conversation we're having. So, in order to properly ignore the possibility we must stop attending to it. One might think this would be difficult, as in the classically impossible order to not think of pink elephants, but Lewis does not seem to appreciate the difficulty. As he says, the possibility can be forgotten because if we pretend that we are ignoring it, we eventually will. He suggests that an aberrant possibility can be forgotten if we "Go off and play backgammon" (560).

It is perhaps not surprising that these responses would underplay the difficulty in switching back to a more credulous context, because they are trying to respond to the skeptical challenge to everyday knowledge, and we do, in fact, tend to be able to forget that we might be a brain in a vat and therefore know *nothing* almost immediately after our epistemological conversation (if we even did truly believe it in the moment rather than just pretending to entertain it). It is less easy to regain lost knowledge when there is a specific challenge to one piece of our knowledge.

Consider the following scenario: your baby is sick, but you know (from reliable processes) that she just has a mild cold and will be fine. When it is pointed out that many quite serious diseases look like the common cold at first, and in this context you are asked what reasons you have to justify your belief that it is merely a common cold, you must admit that (in this new context) you do *not* know that she is only mildly sick. Does this context switch back immediately after a game of Backgammon? It depends. Many factors go into how salient the possibility remains for you, such as your own psychology, the vividness with which the other diseases were described to you, the stakes of the situation (how serious were these other diseases represented to be? How quickly do they take a turn for the worse?), the authority of the person who spoke to you (even though the conversation is the same, we would probably remember this conversation more saliently if it had been with a doctor than with an ever-worried relative), how we are socialized to react to illness, and so on. Switching between contexts is a much more complicated process than contextualist accounts tend to take it to be, and it is inextricably social. (Though we are only discussing the difficulties in switching back to a more credulous context, switching to a doubtful context in the first place also has many complicated social elements such as our trust of the person raising the new possibility and so on.)

In our example of asking for reasons, this is particularly true. Recall that the dominant culture endorses the idea that we ought to be able to provide reasons for our beliefs in order to call them knowledge. Once we are made aware that we cannot provide reasons

in a given case, we may well think that we do not know something we previously would have said we did. Once that conversation is over however, we continue to be unaware of our reasons. If our conviction is particularly strong, we may keep searching about for reasons until we find ones good enough to satisfy us, or we may manage to forget that we do not have good reasons for something through inattention. However, it is at least equally possible that once we realize that we cannot provide reasons for our belief we will no longer believe it, and thus it will not be knowledge for internalists or externalists, despite it being true and our having a reliable process for believing it at our disposal.

Asking for reasons for knowledge can draw our attention to our lack of conscious justification, and cause us to no longer know a given proposition. Though similar to typical versions of contextualism, this loss is not just dependent on conversational context; it is sometimes impossible to regain this knowledge once lost. As we will see in the next section, asking for reasons, trying to answer, and trying to switch back to our naïve state are all inextricably social issues in our epistemology.

IV. Implications

There are many implications and complications not discussed in the above short treatment of asking for reasons. The first is that there is some knowledge which we will not abandon even if we are not able to provide reasons for how we come to know it. What knowledge we are recalcitrant about, and in what circumstances, is just as complicated as the issue of when we are able or unable to switch back to a credulous state, but it might be thought that this recalcitrance gives some comfort to those worried about losing knowledge from questions; if the subject has enough conviction, she will not lose her belief just by engaging in the conversation.

Unfortunately, however, the second implication is that even if the person maintains her knowledge despite not being able to provide reasons, asking her to provide them still has other effects. One of these is that her inability to provide reasons makes it very unlikely that what she's saying will be granted credence by other people that hear the exchange (because the dominant culture accepts asking for reasons as a good test of whether someone has knowledge), resulting in the failure of knowledge uptake by the audience. The other likely effect is that maintaining her belief in the face of being publicly unable to provide reasons will have some mental cost. She will know that she sounds like she is overly credulous, or stupid, or some similarly negative social label, and will also be aware that her hearers are unlikely to now believe what she says. She will also have to

battle her own internal doubts, which even if defeated may well weaken her resolve on an issue that she does, in fact, know.

Relatedly, a third implication is that asking for reasons privileges people who are able to come up with reasons over people who actually have a lot of knowledge. Some people will be able to provide reasons and thus maintain knowledge when others in the same situation would lose knowledge; this is not because of epistemic differences but differences in rhetorical ability. It is perhaps no great revelation to say that people who are able to sound convincing are more likely to have what they say taken up by an audience, and that people who are good at rationalizing things are more likely to be able to convince themselves of things as well. What is perhaps less often considered is that when we ask for someone's reasons, people who are better able to produce convincing-sounding reasons *whether or not they are the actual causes of their belief* will be able to maintain their knowledge, while people who are less able to do this will lose their knowledge. While not quite a Gettier case, it does seem to introduce an unwelcome element, in this case rhetorical skill rather than luck, to having knowledge.

Perhaps the most important implication comes from the fact that we neither ask for reasons every time we hear some new information, nor do we do it arbitrarily. Rather, we tend to ask someone how they know something when what we hear is surprising, or when we do not fully trust the speaker or her knowledge. It is perhaps worrisome that a tendency to ask for reasons for surprising information means that we often cause the speaker to lose knowledge if it does not conform to our worldview, but much more problematic is the tendency to demand reasons from people whom we do not trust. This means that people who are from marginalized groups in our society and who are therefore not trusted to know things will be epistemically oppressed by the increased demand to provide reasons, and their knowledge will in fact have less uptake in the community.

Even worse, if being asked to provide reasons can cause someone to lose the knowledge she previously held, and people from untrusted marginalized groups are asked to give reasons more often, it is possible that the epistemic oppression can extend to forcing that group to have less knowledge than the dominant group.³ This increased questioning is made even worse when we realize that the rhetorical skill discussed

3. Or at any rate this is one possible outcome. There are others. To pick just one example of an alternative, members of the epistemically oppressed group may put more faith in their ability to know things without having to have good reasons for their knowledge—a valuing of “intuition” can be a useful adaptive strategy to epistemic oppression.

earlier that allows one to better maintain and communicate knowledge is not just an individual idiosyncrasy but at least partly the result of socialization, and so people from dominant groups in society not only are questioned less but are better able to defend their knowledge, while people from epistemically oppressed, marginalized groups are both challenged more often and given fewer resources to protect and communicate their knowledge.

How might all this play out? Imagine a parent of a child with some special needs attempting to advocate what she thinks is best for her child to the child's school. If she is constantly doubted and asked to explain how she knows something she testifies to, particularly if the questioner pushes to require deep levels of justification ("How do you know X?" "Because Y." "And how do you know Y?" and so on), this can damage her ability to testify to others and may end up making her doubt herself as well. This is particularly true if she has been socialized to not be as good at that kind of verbal sparring as her interlocutor. All this would make her ability to succeed in this context even more impressive than it already would have been. Demanding she provide reasons for her beliefs is not only an epistemically worrying practice, but a justice issue as well.

V. What is to be Done?

Looking at the implications in section IV, one might wonder what we are supposed to do with this information. Surely we are not supposed to never ask anyone for their reasons again? This seems humanly impossible. Certainly internalists would say it is. As Bonjour says, "The most natural way to justify a belief is by producing a justificatory argument" (1978, 110). Yet, externalists too would be unwilling to give up on asking for reasons altogether. As John Greco says,

It has often been noted that knowledge is a social product with a practical value. We are social, highly interdependent, information-using, information-sharing beings. As such, it is essential to our form of life that we are able to identify good information and good sources of information. In this context, it is not surprising that we make evaluations concerning how beliefs are formed, their history in relation to other beliefs, why they are believed, etc. In other words, it is not surprising that we make evaluations concerning whether beliefs are reliably and responsibly formed. (2005, 266–267)

Presumably one of the best ways we have to make these evaluations is to ask people how they know something. How then to reconcile this with the damage that asking

for reasons can do? Unfortunately, I am unable to offer a perfect solution; it may be the case that some of the very things that make our body of knowledge robust are the *same things* that perpetuate epistemic violence and oppression. That being said, this potentially inevitable damage is no reason not to think about ways to avoid or at least minimize it.

One thing suggested by our discussion so far is that we should be very careful about asking for reasons if we want to avoid eliminating knowledge. Part of being careful means not biasing this demand for justifications toward people from marginalized and epistemically oppressed groups. Making a conscious effort to listen to under-heard voices is at least a good step. Another part of being careful pertains to how we ask for reasons. What strategy is best depends in part on whether one is an internalist or an externalist: lax internalists who truly believe that we ought to be able to have access to our justifications upon sufficient examination should at least make sure that the question does not demand reasons be given immediately or under other kinds of pressure. Externalists who believe that knowledge requires warrant rather than justification should be more careful still; they should ask questions that draw out reliable processes rather than insist on internal access. For an externalist, asking "Have you been right about this kind of thing before?" might be more useful than "How could you possibly know that?". Whether an internalist or an externalist however, another part of being careful is trying to hear the subject's reasons as charitably as possible to try to compensate for the rhetorical advantage some people have over others. As Dotson says in her (2011) article, "*We all need an audience willing and capable of hearing us.*" (239, emphasis in the original)

In the previous paragraph I recommended being careful *if* we want to avoid eliminating knowledge. There are times when this is not the case. One example would be when we have very good *practical* reasons for needing a bias toward false negatives over false positives. For instance, if a friend tells me that she is sure it is safe to jump from a great height into a pool of water below, I may have very good practical reasons for preferring a false negation of her correct belief than a false acceptance of an incorrect belief, and so may have good rational reasons to ask how she comes to know that it is safe. Another example more tied to the justice issues we have been looking at is when members of a marginalized group demand that members of the dominant group provide good reasons for claiming knowledge that supports this relationship of domination. This is both a case where the members of the marginalized group have good practical reason to favor false negatives over false positives, and it may be a case of people from an epistemically oppressed community turning the weapon of doubt against their oppressors. A final example might be as a way to equalize an epistemically imbalanced situation. A teacher might be required to be able to provide better reasons for the knowledge she shares

with her class than her students are, or a doctor might be required to be able to provide better reasons than her patients are, because these roles have so much more *prima facie* epistemic authority. These examples do not abrogate our responsibilities to be careful about asking for reasons, but it does at least provide some situations in which it may be the right thing to do.

Conclusion

In *Sister Outsider* (1984), there is a transcript of a conversation between Adrienne Rich and Audre Lorde. Talking about understanding each other across racial lines despite their shared gender, they have an exchange which I will quote at some length as it is vitally connected to the points I have been making in this paper:

Audre: I've never forgotten the impatience in your voice that time on the telephone, when you said, 'It's not enough to say to me that you intuit it.' Do you remember? I will never forget that. Even at the same time that I understood what you meant, I felt a total wipeout of my modus, my way of perceiving and formulating.

Adrienne: Yes, but it's not a wipeout of your modus. Because I don't think my modus is unintuitive, right? And one of the crosses I've borne all my life is being told that I'm rational, logical, cool – I am not cool, and I'm not rational and logical in that icy sense. But there's a way in which, trying to translate from your experience to mine, I do need to hear chapter and verse from time to time. I'm afraid of it all slipping away into 'Ah, yes, I understand you.' . . . So if I ask for documentation, it's because I take seriously the spaces between us that difference has created, that racism has created. There are times when I simply cannot assume that I know what you know, unless you show me what you mean.

Audre: But I'm used to associating a request for documentation as a questioning of my perceptions, an attempt to devalue what I'm in the process of discovering.

Adrienne: It's not. Help me to perceive what you perceive. That's what I'm trying to say to you.

Audre: But documentation does not help one perceive. At best it only analyzes perception. At worst, it provides a screen by which to avoid concentrating on the core revelation, following it down to how it feels. Again, knowledge and understanding. They can function in concert, but they don't replace each other. But I'm not rejecting your need for documentation. (103–104)

This nicely expresses the tension between our desire to understand what someone is saying and the reasons for their belief on the one hand, and the damage this can do on the other. While presumably Rich was not here asking for reasons because she doubted Lorde so much as because she doubted that she *understood* Lorde, the effect was largely the same. That Adrienne Rich comes from a more dominant group in our society than Audre Lorde does is significant.

Particularly for epistemologists who do not believe that internal access to one's justifications is a requirement for knowledge, it is surprising that asking people how they know something is as widespread as it is, and even for lax internalists it is surprising if they demand reasons at the moment rather than after careful reflection. It is an indication that professional epistemologists accept many of the tenets of our folk epistemology. Asking people for their reasons is a common and perhaps even effective tool for ensuring the value of others' knowledge in our folk epistemological toolkit, but it is also a potent weapon, and we must not ignore this fact.

References

- Alcoff, Linda Martín. 2007. "Epistemologies of Ignorance: Three Types." In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana, 39–57. Albany, NY: State University of New York Press.
- Alston, William. 2005. *Beyond "Justification."* Ithaca, NY: Cornell University Press.
- Audi, Robert. 2001. "An Internalist Theory of Normative Grounds." *Philosophical Topics* 29 (1 & 2): 19–46.
- Bonjour, Laurence. 1978. "Can Empirical Knowledge Have a Foundation?" *American Philosophical Quarterly* 15 (1): 1–13.
- DeRose, Keith. 1995. "Solving the Skeptical Problem." *The Philosophical Review* 104 (1): 1–52.
- Dotson, Kristie. 2011. "Tracking Epistemic Violence, Tracking Epistemic Silencing." *Hypatia: A Journal of Feminist Philosophy* 26 (2): 236–257.
- Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, UK: Oxford University Press.
- Greco, John. 2005. "Justification is Not Internal." In *Contemporary Debates in Epistemology*, edited by Matthias Steup and Ernest Sosa. Hoboken, NJ: Wiley-Blackwell Publishing.
- Lewis, David. 1996. "Elusive Knowledge." *Australasian Journal of Philosophy* 74 (4): 549–567.
- Lorde, Audre. 1984. *Sister Outsider*. New York, NY: Ten Speed Press.
- McDowell, John. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.
- Sosa, Ernest, Jaegwon Kim, Jeremy Fantl, Matthew McGrath. 2008. *Epistemology: an Anthology*. Malden, MA: Blackwell Publishing.
- Townley, Cynthia. 2011. *A Defense of Ignorance: Its Value for Knowers and Roles in Feminist and Social Epistemologies*. Lanham MD: Rowman & Littlefield Publishers.
- Tuana, Nancy & Shannon Sullivan. 2006. "Introduction: Feminist Epistemologies of Ignorance." *Hypatia: A Journal of Feminist Philosophy* 21 (3): 485–509.

Journal of Cognition and Neuroethics

Brain Rays, Advertising, and Fancy Suits: The Ethics of Mind Control

Brent Kiou

University of Utah

Biography

Dr. Brent Kiou is currently a psychiatry resident at the University of Utah. He completed medical school at the David Geffen School of Medicine at UCLA, and did his doctoral work in philosophy at UCLA. His research interests include applied ethics, metaethics, contract theory, and philosophy of psychiatry, and he has published on the ethics of enhancement, genetic discrimination, and the relationship between decision-making autonomy and values.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2014. Volume 2, Issue 1.

Citation

Kiou, Brent. 2014. "Brain Rays, Advertising, and Fancy Suits: The Ethics of Mind Control." *Journal of Cognition and Neuroethics* 2 (1): 191–210.

Brain Rays, Advertising, and Fancy Suits: The Ethics of Mind Control

Brent Kious

Abstract

Many science-fictional kinds of mind control—technologies that might be used to manipulate other persons' thoughts and behavior—seem patently impermissible. The most natural account of this impermissibility is Kantian: mind control technologies would undermine others' rational capacities, but our duty to respect each others' rational personhood precludes this. I argue, however, that the Kantian view is inconsistent with the fact, demonstrated by several examples, that we permissibly manipulate others' thoughts and behavior in a variety of mundane ways on a regular basis. After considering possible defenses of the Kantian view, I delineate an alternative theory, according to which we can distinguish permissible from impermissible sorts of mind control on largely contractualist grounds.

Keywords

Kant, rationality, mind control, contractualism

Introduction

Mind control sounds scary. The thought that a person might use brain rays, psychosurgery, or some novel technological process to make others think and do things horrifies us. Here I will assume that this horror reflects a moral truth: the use of certain sorts of mind control—like a brain ray or a neural parasite or any of the other products of science-fiction writers' fevered imaginations—would be terribly wrong. In what follows, I will try to make sense of this fact. Though the project is more than a little whimsical, it has three serious motivations. First, and most obviously, there is the motive of applied ethics: sophisticated mind control methods may someday become reality, and we need to know what to say about them. Second, as a psychiatrist, I am interested in the many criticisms leveled at my field—one of which is that psychiatrists often wrongly control others' minds—and I hope that thinking about very extreme sorts of mind control may shed light on those controversies. Finally, there is the motive of philosophical ethics, and in specific the matters of examining the ethics of consent and evaluating one common interpretation of the Kantian notion that we should respect others' rational personhood. This latter principle seems to provide the most natural account of why mind control would be wrong, so thinking about mind control may help us examine it.

In Section 1, I start by defining the term *mind control*, and in the process say a little about what I take rationality to be. I then briefly review the Kantian account, outline how it could explain the wrongness of things like brain rays, and compare it to similar positions taken about other sorts of manipulation to which persons may be subjected, such as lying. In Section 2, I point out that, despite the seeming impermissibility of mind control, we permissibly control others' minds all the time, in ways that are hard to differentiate from clearly impermissible sorts of mind control given only the resources of the Kantian view, and which cast that view into doubt. In Section 3, I consider some possible rejoinders to the examples used in Section 2 and the criticism of the Kantian view that follows from them. Finally, in Section 4, I offer the beginnings of an alternative theory, suggesting that a loosely contractualist approach may succeed in explaining both when mind control is permissible and when it is impermissible.

1. Background and Definitions

Throughout what follows, I use the term *mind control* to refer to what we might also call *anti-rational influence*, where this is causing someone to think or do something irrespective of whether she has adequate reason to think or do it, in a fashion that bypasses or undermines her rational capacities. By *think something* I mean *form or abandon a judgment-sensitive attitude* (Hieronymi 2006), where judgment-sensitive attitudes include beliefs, desires, and intentions.

The notion of rationality assumed in this paper is merely a formal one. That is, I want to assume only that being rational requires thinking and acting in the ways one has adequate reason to think and act, without taking a stance on what those reasons are. Put in terms of judgment-sensitive attitudes, we might say that being rational is a matter of adopting, through the exercise of one's rational capacities, those beliefs, desires, and intentions one has adequate reason to adopt. Being irrational is, on this view, a matter of failing to think or act in the ways one has adequate reason to act.

In what follows, I will not offer an account of the nature of reasons, taking this to be a ground-level notion. I do, though, want to observe that any plausible notion of rationality needs to assume that reasons are usually internal (Williams 1981), since we typically do not want to say that a person has been irrational just because her judgment-sensitive attitudes fail to reflect reasons which were not available to her anyway—ignorance and irrationality are different matters. Still, our notion of rationality also needs to incorporate standards of sensitivity to external reasons. Rationality is not merely a matter of thinking and acting according to the reasons one *takes* there to be, but is

also a matter of responding to reasons that there are. Although it would not necessarily make one irrational to have a false belief if the error is due to limited information, it might make one irrational to have a false belief if the error occurred because one was not appropriately attuned to evidence that contradicted it.

I want to avoid giving an account of what reasons we have—that is, to avoid giving a substantive account of rationality—for two interrelated reasons. First, which theory of practical rationality is correct is controversial: one could claim, for example, that to be rational is to maximize one's own utility or happiness, or instead that satisfaction of one's actual or even merely perceived interests is rational. Second, for present purposes, no opinion about what is substantively rational is necessary, as the anti-rational influences we will examine can be assumed to interfere in the *processes* of reasoning as well as in their practical consequences. Moreover, the Kantian view about rationality which will be my target is itself relatively silent about substantive matters.

Although generally anti-rational influences cause someone to think or act in ways she does not have adequate reason to think or act, they come in two different types.¹ First, there are *anti-rational influences properly considered*, where a person is influenced, by a mechanism that bypasses or undermines her rational capacities, to do something she would not have done, or chosen to do, otherwise—as when X uses a brain ray to cause Y to give X all of his money, but Y never would have thought to do so without that influence. Second are what we might term *para-rational influences*. In these cases, a person is influenced to do something she would have done anyway, but the influence still undermines or bypasses her rational capacities. For instance, an insurance salesman might motivate me to buy life insurance by frightening me with ghastly tales of what could happen to my family without it, where that fear undermines my rationality, even though I would also have bought the life insurance if I were given the opportunity to reason about the decision. In that I suppose it is harder to defend the permissibility of anti-rational influences properly considered, I will focus on the former, and assume that what holds for the anti-rational properly considered also holds for the merely para-rational.

Finally, it is important to point out that anti-rational influences, even properly considered, vary with respect to how dramatically they shift a person's attitudes and actions. One might influence someone to do something she would not have done otherwise but still induce only a small shift in her inclinations, as when Y is undecided between A and B, though leaning slightly more strongly toward B, but X persuades him to do A by giving him a mind control serum that makes him suggestible and then telling

1. Thanks to Jesse Summers for pointing out this distinction.

him to do A. Para-rational influences can also vary in this way, as when Y is more inclined to do B than A, but X nevertheless cements his decision by giving him the mind control serum.

Again, I assume that most of us regard very extreme, science-fictional sorts of mind control as quite clearly wrong. Joel Feinberg artfully expressed this sentiment when, in surveying the moral valence of all varieties of human oppression, he wrote:

[M]ost odious... is the manipulation of a person without his consent. Patients or prisoners ... can be drugged, put under total anesthesia, and then made to undergo lobotomies or other kinds of surgical manipulation or mutilation of the brain. Psychotropic drugs used in small quantities and electric stimulation of the brain for short periods have less severe effects and are revocable, but when imposed on a person without his consent ... they are hardly distinguishable on moral grounds from assault and battery. (Feinberg 1987, 67)

Feinberg himself is not much of a Kantian, but I suspect that to most modern philosophers the most natural account of the wrongness of things like brain rays and other technologically advanced types of mind control is basically a Kantian one. Kant famously claimed that we must always treat others' humanity never merely as a means, but also as an end in itself (Kant 2002). For Kantians, *humanity* is often equated with *rational personhood*, and Kant's admonition is understood to mean that we should never manipulate, undermine, corrupt, or exploit others' rational faculties. One might think that this is what makes lying wrong—by lying, I exploit your rational faculties to get you to do something—and it also seems to explain why it would be wrong to use a brain ray to force a person to think or act a certain way: in so doing, one would make him think something in an anti-rational fashion, thereby undermining or exploiting his rational faculties and showing disrespect for his humanity. For Kantians, this is not only a wrong, but the most fundamental of wrongs.

Though I am not aware of anyone who has made this particular argument about mind control in the philosophical literature, related positions abound. For instance, Ginger Hoffman has recently argued that it is morally wrong to take antidepressants without engaging in psychotherapy because it *objectifies* oneself—that is, runs roughshod over one's rational personhood—by making one think things without engaging one's reason (Hoffman 2013). The Kantian stance also underpins much of the literature on coercion and undue influence, and is widely regarded as the basis for the doctrine of informed consent in bioethics (Beauchamp and Childress 2001). There, it is commonly supposed that what

makes consent essential is our obligation to respect another person's rational autonomy (his capacity to govern his own life through the exercise of his rational capacities), and it is only when an agent lacks our usual rational capacities (because he is developmentally disabled or mentally ill or already under the control of someone else) that we need not respect his decisions about his life.

2. Permissible Mind Control

Accordingly, I take it that the Kantian account of the wrongness of mind control is natural and appealing, even if it has not been widely defended. I also think it has some worrisome inadequacies, at least in the relatively simple form presented above. In particular, we can show that rationality need not always be respected: ordinary life abounds with cases where it is *permissible* to make others believe or do things without giving them good reasons, thereby undermining or circumventing their rationality. For example, consider the following case:

Connie Consumer wants to buy a watch. She is in a department store, contemplating both a cheap digital watch and a Rolex. Suppose she is more inclined, initially, to buy the cheap digital watch. Sally Salesperson thinks Connie is better off with the Rolex; she can tell by the way that Connie is dressed that she is wealthy, and it is a better watch for the money. So Sally persuades Connie to buy the Rolex. She does this in part by offering Connie reasons for buying it. But she does other things, too: she puts the Rolex under a bright light on a black background, so it sparkles appealingly. She draws Connie's attention to a nearby advertisement depicting a beautiful woman strolling down the Champs Elysée wearing the same watch. She models the Rolex for Connie, handling it reverently, but treats the digital watch as though it were slightly soiled. Ultimately, Connie buys the Rolex.

I assume that what Sally does in this case is permissible: she is just engaging in ordinary sales tactics, and these are a widely tolerated practice about which we have few, if any, moral qualms. But I also think some of Sally's actions amount to anti-rational influence, in that she gets Connie to (want to) do something without offering her reasons to (want to) do it. It seems plausible, for instance, that Connie is influenced, at least marginally, by the poster displayed behind the display counter. But when Connie decides to buy the Rolex, she does not think: "oh, that's a really nice poster there, with that beautiful model wearing this watch. *That's* a reason for me to buy it." Or, if she did think anything like

that, she would be irrational (or at least mistaken), since she would be taking herself to have a reason she does not. Rather, the poster presumably increases the likelihood that Connie will buy the Rolex by causing her to associate the watch with romance, exotic settings, and other things that she desires—even though having the Rolex typically would not make the satisfaction of those other desires more likely.

Likewise, the facts that the Rolex sparkles appealingly beneath the display lights and looks good on Sally's wrist are not reasons for Connie to buy it. They may highlight some reasons to buy it, namely the facts that the watch is beautiful and that it would look attractive on her own wrist. But Connie is probably already aware of these reasons. If placing the Rolex under the lights and modeling it affect her decision, they do so either by causing her to pick the Rolex without deliberation, or else changes the salience such reasons have for her, perhaps by making her attend to them more fully.² Either way, these influences seem to subvert her rationality, by making her decision depend in part on factors outside of reason.

Finally, the different ways Sally treats the Rolex versus the cheap digital watch might affect Connie's decision in an anti-rational fashion. Obviously, one reason to buy a Rolex instead of a cheap digital watch is that Rolexes have a certain cachet: they increase your social status, at least marginally. Sally's ability to influence Connie's decision by handling the watches differently plays upon this reason, but presumably is neither itself a reason (that *Sally* will think more highly of Connie if Connie buys the Rolex is generally not, absent some special story, a good reason for Connie to buy it), nor even reminds Connie of the reasons that she has (phenomenologically, it seems implausible that Connie recognizes what Sally's different treatments of the watches signify about their effects on status, and then chooses the Rolex on the basis of its potency as a status-symbol). Rather, as anyone who has been influenced in this way may be able to attest, it is something like the triggering of her instinctive desire to achieve social status or to display its trappings that allows Sally to influence Connie in this manner.

Sally's ability to influence Connie is just one of many examples furnished by modern consumer marketing, where permissible mind control techniques abound. Obviously, print, television, and internet advertisers have caught onto the fact that putting scantily-

2. It is well known that even the physical proximity of a desired object can influence persons' decision-making. For instance, Bushong et al. (2010) showed that subjects' willingness to pay for snack-food items (a measure of how much they want those items) is significantly higher if those items are physically present rather than simply displayed on a computer screen, even if they will be delivered in the same time-frame. It is also widely accepted that drug addicts are more likely to relapse if they see the drug of abuse or are given other reminders of their use (Robbins and Everitt 1999).

clad persons in their advertisements can improve the sales of almost anything. This strategy may be unethical because it exploits the models, but the fact that it slightly compromises consumers' rationality does not seem to make it so. Celebrity endorsements have the same effect (Spry et al. 2011). The fact that a popular NASCAR driver appears on television endorsing motor oil he is paid to endorse is no reason whatsoever for me to buy that motor oil, but may influence me even so. Again, however, this influence seems morally permissible. Physical stores use anti-rational influences, too. In addition to the techniques of salespersons, grocery stores will do things like locating bread, cheese, milk, and meat near the rear of a store, so we have to wander long aisles full of things we were not initially planning to buy to make our planned purchases; likewise, many groceries contain in-house bakeries that make small quantities of fresh bread throughout the day, often at a loss, because the smell of fresh bread is comforting and makes people hungry, inclining them to buy more in the rest of the store (*Economist* 2008).

Nor are anti-rational influences restricted to the commercial world. They occur in other aspects of social life, too. It is clear our physical appearance can affect others' thoughts about us in subtle ways. We all know that our relative pulchritude can alter how others treat us—how likely they are to bestow favor upon us, how they assess the merits of our ideas, and how much authority they cede to us. The psychological literature bears this out (Baker and Churchill 1977). Likewise, we frequently take pains to exploit this fact: we might wear a fancy suit to a job interview, or submit an attractive photograph with a job application, with the clear aim of increasing our chances of success. I take it that, on the whole, these efforts are permissible, but I also suppose their effect is often anti-rational. While one might argue that personal hygiene and grooming are moderately reliable indicators of the qualities that employers usually desire, the attractiveness of a persons' face is probably not (unless the employee is being hired for a position where attractiveness is itself essential).

For a somewhat different example, note that I sometimes wear glasses when I am giving lectures or presentations. I could wear contact lenses instead. But suppose I were to choose to wear glasses on some occasions because I know from having read the psychological literature that they subtly but measurably improve my audiences' implicit assessments of my intelligence, thereby (presumably) increasing the likelihood they will conclude my arguments are sound (Thornton 1994; Manz and Lueck 1978). I think most of us would regard my wearing glasses for this reason to be morally permissible. But again, it probably represents an anti-rational influence: my glasses do not provide anyone with a *reason* to conclude my arguments are sound: people who wear glasses are not more likely to make sound arguments.

Some things we do to influence others in anti-rational ways quite closely resemble habits and reflexes and may be hard to distinguish from them. Facial expressions, posture, and gestures all affect others' thoughts about us and their responses to what we say and do, but generally do so without giving reasons. As a psychiatrist, I sometimes try to look more sad and worried for my patients than I may actually feel in the moment. Although I invariably care about them and want them to get well, occasionally it is hard to muster genuine empathy for them; one's emotions can become fatigued. Fortunately, after some practice, I know how to make myself look and sound like I feel sad and dismayed even when I do not feel that way. I also know that getting my patients to believe I am sad or dismayed can make a difference to their treatment: it can make them more likely to reveal important information and more likely to follow my recommendations about their care. Admittedly, it could be somewhat discomfiting to them to learn I do this, but I do not suppose the practice is unethical. On the contrary, mustering affective displays that are somewhat insincere seems required both by my concern for my patients' well-being, as well as by my professional (and ultimately, moral) obligation to treat them as well as I am able. But once again, it seems clear that influencing my patients through these artificial displays of emotion represents an anti-rational influence: I manage, I suppose, to make them have beliefs (as well as feelings) about me that make them more likely to do things I think they ought to do, without giving them genuine reasons to do those things (though clearly I assume they have other reasons to do those things).

In sum, then, I take it that these examples strongly suggest that anti-rational influences—including both para-rational influences and anti-rational influences properly considered—can be morally permissible. But if that is the case, it follows that it is sometimes permissible to intentionally compromise others' rationality, and the Kantian's claim that we have an obligation to respect others' rational personhood, where this is taken to entail that we must refrain from doing things that compromise their rationality, must be false. This leaves us unable to explain why mind control techniques that seem obviously impermissible are so.

3. Objections and Responses

Of course, proponents of the Kantian position have a variety of responses to the foregoing argument at their disposal, which it should be worthwhile to explore here.

First, one might simply insist that all the sorts of "permissible" mind control described above are really impermissible. This insistence is not unmotivated. After all, it has been argued before that advertising with a persuasive (as opposed to merely informative)

focus is impermissible (Santilli 1983; Crisp 1987) and one can imagine taking the stance that advertising in general—prone as it is to making us want things we do not need, leading to increased dissatisfaction or, alternatively, to run-away consumption, the waste of precious resources, environmental degradation, and so forth—ought to be prohibited (Carlson et al. 1985). This is not exactly a Kantian motivation, but we can also concede that advertising looks more distasteful the more closely we scrutinize it, that certain sorts of advertising techniques (e.g., subliminal advertising) are clearly impermissible, and that the more effective advertising is in motivating a person to purchase an item—especially when the influence is properly-considered anti-rational—the less permissible it seems.

Even admitting these other considerations, however, I think we should not follow the Kantians in drawing this hard line. Though the matter comes close to being a fundamental disagreement, we should note two things. First, the Kantian position would commit us to a revisionist stance regarding a large number of practices that now appear to be morally permissible—such as dressing nicely to improve our prospects of getting a job, wearing makeup, mirroring a patient’s emotions in a therapy session, or even ordinary sales tactics. This seems to me to be enough reason to reject the Kantian view. Second, we should note that this is really just one species in a larger genus of problems for Kantians, so that a revisionist stance with respect to these cases would commit us to revisionist stances elsewhere. In particular, the standard Kantian admonition to respect rational personhood is also taken to apply to our *own* rational personhood, and would seem to exclude the use of alcohol and other mind-altering substances that many of us regard as at least occasionally permissible. Conversely, if we are willing to make exceptions in this sort of case, then we should be more willing to admit exceptions in cases of mind control.

One might also object that the plausibility of the cases above hinges on an equivocation between anti-rational influences properly speaking and para-rational influences. According to this objection, the examples of mind control above seem permissible largely insofar as they are para-rational, but para-rational influences need not be regarded by Kantians as impermissible. The motivation for this objection is that para-rational influences in general seem less problematic than anti-rational influences properly speaking. For instance, it is clearly permissible for a commercial to be used to motivate me to purchase a new car if I am already inclined to purchase a new car (and have sufficient reasons to do so), but it is less clearly permissible for someone to use a commercial to motivate me to buy a car when I had no such inclination antecedently or, worse, was inclined *against* buying a car.

Still, this objection rests on two mistakes. First, it is false that all of the cases above are correctly interpreted as (merely) para-rational and not as anti-rational properly speaking. In the case of Connie Consumer, we assume at the outset that Connie is inclined to buy

the cheap digital watch instead of the Rolex. Nor is this assumption implausible. More generally, the net effect of things like consumer marketing techniques is not, overall, to get people to do things they were going to do anyway, but to do things they were not going to do otherwise (it would otherwise not be economically efficient). One supposes that sales techniques can increase a consumer's willingness to pay, the sales of a particular item, or the overall consumption of a particular commodity—all of which would be best explained by a capacity to influence others to do things they would not have done in the absence of the influence.

Second, it is not at all clear that Kantians should be friendlier toward para-rational influences than they are toward properly-speaking anti-rational influences (that is why I have persisted in placing them in the same category, anti-rational influences generally). The Kantians' position looks to the mechanism of an influence, rather than its outcome—what matters is not *what* an agent decides to do as the result of our interactions with him, but *how* we brought him to decide to do it, and different influences may undermine a person's rationality to the same extent even if one is para-rational and the other properly anti-rational. Indeed, the Kantian requires a process-oriented as opposed to outcome-oriented position about what makes mind control wrong in order to make sense of certain intuitions, such as that it would be wrong to use a brain ray to manipulate a person into doing something he would choose to do anyway. The Kantian takes the same position about coercion: that it is wrong to coerce someone into doing something even if he has a reason to do it and even if he would decide to do it on his own, because coercion subverts his rational personhood and thus his autonomy.

So this second objection seems to fail. There are, however, other ways to reinterpret the cases. For instance, many Kantians accept that we can make it permissible for others to do things to us that would usually be impermissible if we consent to that treatment in advance. Often, implicit or tacit consent is adequate for this purpose. Thus, one might claim that what makes Sally's manipulation of Connie, or advertisers' manipulation of consumers generally, or my influence on my patients—is that in each case the parties influenced know that they are going to be influenced before proceeding, and by proceeding give tacit consent to that influence. The plausibility of this objection is suggested by a revision of Connie's case: imagine that instead of acting like the salesperson she is, Sally masquerades as another shopper but still does many of the things she did in the original example to influence Connie's decision to buy the Rolex. Her role here seems more deceptive, and correspondingly more worrisome—presumably because Connie *did not* implicitly consent to being influenced in this way.

While we can concede that consent has a transformative moral force, this objection only passes muster if we assume that persons who enter a store, or watch television advertisements, or who come into a psychiatrist's office for a consultation, *know* that they are likely to be subjected to the influences outlined above. But that seems implausible, both because this does not seem to be the sort of thing of which people are typically cognizant, and because many of these influences would, one supposes, be rendered ineffective if we were aware of them. This certainly seems true of my patients' responses to my emotional displays, and (though I am aware of no psychological data that supports this claim) seems likely to be true of sales techniques and advertising.

For a fourth objection, one might be tempted to explain the influences in the cases above as permissible because they ultimately *enhance* our rationality or are parasitic on processes that are essentially rational. For instance, suppose Sally partly brought about Connie's decision to buy the Rolex just by taking it out of the display case, increasing its physical proximity to Connie. One could claim that this influence depends on a rational (albeit implicit) judgment on Connie's part, one that reflects discounted utility theory. People tend to value rewards that are more temporally proximate over equivalent, and even larger, rewards that are temporally distant (Ainslie 1975). Discounted utility theory claims that this is rational: we *should* discount rewards at a rate that is proportionate to their temporal distance, since the more distant in time a reward is, the less likely it is that we should ever receive it. Similarly, one might claim that Connie's adjusting her relative valuation of the Rolex when it became more physically proximate reflects an underlying decision-making mechanism that embodies a rational principle, according to which rewards are discounted more the farther away from us they are.

This objection also fails, however. First, even assuming it is sometimes rational to discount utility by physical proximity, this seems like a spurious rationalization for Connie's decision, since in this particular case the physical proximity of the Rolex says little about the likelihood that Connie will be able to obtain it (that depends mostly on how large her credit card limit is). More generally, showing that our decisions sometimes invoke a mechanism that is rational as a general strategy or heuristic or which may have been evolutionarily advantageous does not make those decisions rational themselves. Finally, the literature on temporal discounting suggests that we do not always respect discounted utility theory: decision-makers do not apply their discounts in a consistent fashion, exhibiting what is known as dynamic inconsistency. We will often, for instance, value reward A over the larger reward B when A is immediate and B is temporally removed by time T1, but will value B over A when A is delayed by time T2 and B is delayed by T1 + T2, so that we have discounted B by proportionately more than A in the first case but

not the second (Kirby and Herrnstein 1996). This violates discounted utility theory, and implies that the underlying structure of our decisions is, in this respect, irrational.

4. An alternative account

If you take any of the examples I gave above as cases of permissible mind control, then one has to conclude the Kantian position I outlined earlier is inadequate, because it is sometimes permissible to do things to others that undermine or exploit their rational faculties. Now I want to offer an alternative to the Kantian position. The alternative has three elements, and aims both to explain why things like brain rays would be wrong and why mundane sorts of mind control are often permissible.

First, let me define what I'll call the relative *violence* represented by a particular mind control technique. Violence, in this context, is a composite of two other properties, namely *power* and *invasiveness*. Power is the propensity of a technique, considered broadly across an array of possible settings, to induce persons to form judgment-sensitive attitudes in a way that does not reflect the balance of reasons. Wearing glasses during a lecture is not a particularly powerful influence: it is unlikely to persuade you that I am making good arguments if I am not, for example. In contrast, a brain ray or a well-told lie can be quite powerful, in that it is (by hypothesis) able to get people to form attitudes contrary to the reasons they have.

Invasiveness is related to power but importantly different. It is a technique's capacity to affect aspects of a person's mind or thinking that are more or less central to his identity, irrespective of the balance of reasons. Suppose there is a brain ray that can affect persons' attitudes only when the balance of reasons is equivocal, so it is not powerful in the sense outlined above. Now assume that I am a Zoroastrian and someone uses the brain ray to turn me into a Buddhist. Even supposing I had no reasons to accept Zoroastrianism over Buddhism, the forced change in my religious beliefs seems extremely objectionable—and thus, violent in my sense of that term. Presumably, what makes it violent is that it changed a characteristic essential to my identity.³

3. This example raises an interesting question: since most religious conversions presumably do not hinge on the balance of reasons, but almost always affect core components of identity, shouldn't it nearly always be impermissible to try to convert someone? Most of us are inclined to think that it is not, but that raises a question about whether invasiveness really matters at all. I am tempted to say that standard religious conversions actually are not invasive in the sense I intended because, even though they change features of a person's beliefs that are central to his identity, they do so in an identity-preserving fashion: that is, they depend upon the exercise of his rational faculties and typically this results in a set of beliefs with which the agent actively *identifies*. But this response fails, since we can imagine a brain ray that would work in more or less the same manner, and which

My second observation is that the permissibility of any particular mind control technique reflects, to some degree, whether its relative violence is sufficiently counterbalanced by the expected risks and benefits to all concerned parties, compared to alternative courses of action. Overall, it seems like even extremely violent mind control techniques can permissibly be used in extreme circumstances, and that even relatively non-violent techniques should not be used if the expected harms are large and the expected gains are small. Together, these factors account for some of the following intuitions:

- a. It is often wrong to use especially violent mind control techniques even if the aim is beneficent. For example, I ought not use telepathy to make a patient take his antihypertensive medications, even if it would be better for him, and even if no other methods of persuasion exist.
- b. It can be permissible to use even very violent techniques of mind control if the benefit/harm balance is sufficiently large. For instance, Professor X can permissibly use his telepathic powers to stop Alice from shooting someone or crashing her car into a wall.
- c. It is often impermissible to use even non-violent mind-control techniques when the expected harms are very large and not outweighed by the associated benefits: for instance, it would be wrong for Marilyn Monroe to use her feminine wiles to persuade President Kennedy to give her nuclear access codes so she could sell them for a profit.
- d. Finally, others can permissibly treat us in anti-rational ways to benefit themselves, even if there are costs to us, as long as the benefits to them tend to outweigh our costs and the treatment is not too violent; examples include when the grocery store owner or advertiser or car salesman manipulates us into buying things we might be better off without or into spending more money than we might otherwise do, or an interviewee's charming smile persuades us to choose him instead of a more qualified candidate.

would still seem impermissible.

A different factor we should consider is that in the standard cases, a person is (I assume) usually converted by a true believer. This seems to be important to the permissibility of the conversion, in that it would be morally very different to convert someone to a set of religious beliefs one does not accept oneself (whether that be for one's amusement or for material gain). In forming our intuitions about conversion-by-brain-ray cases, we may be assuming that the brain-ray operator does not accept the beliefs to which he converts others.

But in the end, I think a contractualist account is essential to differentiating these cases; for more on that, see below.

I should emphasize, however, that I do not suppose—or at least am not myself able to identify—anything like a formula for balancing the violence of a mind control technique with its risks and benefits. Nor do I even suppose that the loosely consequentialist stance outlined above amounts to the whole story. It is merely one set of considerations among potentially many others that inform our judgments about what is permissible.

This brings us to the third element of my proposal. I think we can understand the way we balance risks, benefits, and violence when determining the permissibility of mind control in a roughly contractualist way. I do not have the space to delve into the details of contractualism here, nor to defend contractualism as a moral theory. It should be sufficient to remark that according to most contractualist moral theories, actions are impermissible if and only if it would be reasonable for us to choose, under unbiased conditions, to prohibit them (Scanlon 1996). So consider clearly impermissible sorts of mind control first. Plausibly, we could all reasonably agree, if we were impartial, not to subject each other to extremely violent methods of mind control, at least unless the situation were dire. This is particularly obvious when such methods would impose severe harms on us to only slightly benefit others—mind control in those cases would be wrong for the same reasons that theft and violence often are. Moreover, it is reasonable for us to be disposed against especially violent mind control techniques even when the net benefit of their use in a particular case is great, because the temptation to utilize such techniques inappropriately is high: it behooves us to assume there is a standing prohibition against such influences even if that prohibition is not absolute. Something similar is true of coercion: although it can be permissible to coerce others when the balance of risks and benefits is clearly favorable (as when we do so to prevent violence against others or to encourage them to adopt behaviors, like wearing seatbelts, that have great benefit but minimal cost), we do well to use coercion only with great caution.

On the other hand, given a contractualist background, we can explain why certain mundane, non-violent mind control techniques are permissible. This is because they represent, or are at least closely *related to*, essential aspects of human social practices which could not feasibly be eliminated from our lives, either because they are so often beneficial to us collectively, or because any efforts to prohibit them in some cases would alter other aspects of our lives undesirably. Given such features, it would not be reasonable to choose together to prohibit them *even though* they compromise our rationality. We want to be allowed to smile at people even when we do not feel happy, to dress nicely on important occasions even if this may be misleading, and possibly, to wear glasses when we give philosophy lectures, because these types of interventions make social life better, or are at least practically indistinguishable from other actions that do. And, I suspect, the value

we attach to our ability to do these things is great enough that we are willing to tolerate others' occasionally performing similar actions in ways that undermine our rationality. Even if, in a more ideal world, we could interact with each other on a purely rational level, our capacities to influence others' thinking in anti-rational ways, and perhaps even our capacity to be so influenced, are desirable features of our embodied personhood. If the alternative to being occasionally gamed and tricked and confounded by others is adopting a hypersensitivity to anti-rational influences that would dramatically alter the sorts of creatures that we are, we should clearly choose the former.

The contractualist reasoning offered above obviously will not hold for all sorts of permissible anti-rational influence. For instance, it does not seem to apply to persuasive advertising techniques, which are not, so far as I know, integral to human social life. But I suspect that in many of these cases other contractualist accounts are available. In the case of advertising, one might suppose that it is rational for us to allow persuasive advertising despite its potential for anti-rational influence for something like the following reasons: (1) most of us greatly value having access to print, television, radio, and online media; (2) the production and distribution of these media are costly and in general those costs cannot be covered by subscription fees; (3) persuasive advertising that is itself visually captivating or entertaining but which also has the potential for influencing consumers in anti-rational ways is an efficient method for covering media costs; (4) alternative sorts of advertising that are purely informative and accordingly bland and uninteresting, but which also would not influence anyone in anti-rational ways, are not efficient ways of covering media costs (because they are less effective than persuasive advertising at capturing consumers' attention and affecting their purchases); (5) in the main, the extra costs imposed on us by advertising are sufficiently counterbalanced by the benefits of access to media; and, (6) persons who do not wish to be influenced by persuasive advertising can simply limit their exposure to the media that contain it. Of course, whether these considerations are sufficient to render persuasive advertising permissible depends on whether each of considerations (1)-(6) above is true, which is mainly an empirical matter. My purpose here is not to make an argument for the permissibility of advertising *per se*, but just to illustrate the sorts of considerations such an argument might include.

Applying a loosely contractualist way of thinking to mind control has useful implications. For instance, it is a way of making sense of—and partially⁴ justifying—the

4. Only partially, however, because it is clearly possible that some of our moral reactions to novel technology are simply unfounded or irrational.

tendency many of us have to regard technologies as impermissible just because they are new and unfamiliar. For most of us it would seem *more* problematic to increase others' estimations of our intelligence by using a brain ray than it would be to increase their estimations of our intelligence by wearing glasses, even if we assume the power and the invasiveness of the two techniques are equivalent (i.e., that this is all the brain ray in question can do). This seeming discrepancy can sometimes make sense from a contractualist viewpoint, for two reasons. First, wearing glasses to impress others with our intellectual ability bears a significant resemblance to other things we reasonably want to be free to do, such that it might not be reasonable for us to prohibit the practice. This does not hold for using a brain ray to induce similar beliefs. Second, at least given the way the case above is described, we have little information about the different ways the brain ray technology in question could be developed or applied, and thus little information about the other sorts of manipulation it could make possible. Although to some extent we could assess each of these future developments on its own merits as it occurred, we also recognize that social discourse about the permissibility of novel practices (not to mention actual legislation) often lags far behind practices themselves (this is perhaps one of the main motivations for bioethics: that we should think about the moral implications of new technologies before we use them, because it is often hard to go back). Thus, permitting the use of a novel technology might seem unreasonable, even if the potential harms of the technology are not significantly different than those of practices with which we are already familiar, just because we cannot reliably, either theoretically or in practice, differentiate that technology from others it would clearly be reasonable to prohibit.

5. Conclusions

The foregoing arguments suggested that the Kantian contention that our rational nature should be inviolable is inconsistent with how we usually live our lives, wherein we permissibly utilize all manner of anti-rational influences. This implied that Kantians are unable to account for why some mind control techniques are impermissible, paving the way for an alternative theory. To be sure, there is nevertheless much that is correct about the Kantian view—our rationality is undeniably a good, and a distinctively human good at that. Indeed, one wonders whether the best way of accommodating Kant's recommendation to respect others' rational personhood does not involve something along contractualist lines. Contractualism has its roots in Kant's ethics, and many contractualists (such as Scanlon) think that our obligation to abide by rules that we would reasonably choose together is underpinned by an obligation to respect each other as rational. It

would of course, be ironic if it turned out that respect for rationality sometimes permitted undermining rationality, but this seems consistent with much of our experience. In the end we are not *ideally* rational beings, and we are not *merely* rational beings: much of what we do, and much of what we think, is the product of irrational, or at least non-rational, aspects of our selves. Although our rationality is valuable, it is also sometimes *rational* to trade it against other considerations that have value; we are complicated and messy and emotional, and it is sometimes best to embrace these facts about ourselves.

References

- Ainslie George. 1975. "Specious reward: a behavioral theory of impulsiveness and impulse control." *Psychological Bulletin* 82 (4): 463–496.
- Baker, Michael and Gilbert A. Churchill Jr. 1977. "The impact of physically attractive models on advertising evaluations." *Journal of Marketing Research* 14 (4): 538–555.
- Beauchamp, Thomas L. & James F. Childress. 2001. *Principles of biomedical ethics*. New York: Oxford University Press.
- Bushong, Benjamin, Lindsay M. King, Colin F. Camerer, and Antonio Rangel. 2010. "Pavlovian processes in consumer choice: The physical presence of a good increases willingness-to-pay." *The American Economic Review* 100 (4): 1556–1571.
- Carson, Thomas L., Richard E. Wokutch, and James E. Cox Jr. 1985. "An ethical analysis of deception in advertising." *Journal of Business Ethics* 4 (2): 93–104.
- Crisp, Roger. 1987. "Persuasive advertising, autonomy, and the creation of desire." *Journal of Business Ethics* 6 (5): 413–418.
- Feinberg, J. 1987. *Harm to Others. The Moral Limits of the Criminal Law. Volume I*. New York: Oxford University Press.
- Hart, Joshua, James A. Schwabach, and Sheldon Solomon. 2010. "Going for broke: Mortality salience increases risky decision making on the Iowa gambling task." *British Journal of Social Psychology* 49 (2): 425–432.
- Hieronymi, Pamela. 2006. "Controlling attitudes." *Pacific Philosophical Quarterly* 87 (1): 45–74.
- Hoffman, Ginger A. 2013. "Treating Yourself as an Object: Self-Objectification and the Ethical Dimensions of Antidepressant Use." *Neuroethics* 6 (1): 1–14.
- Kant, Immanuel. 2002. *Groundwork for the Metaphysics of Morals*. New Haven: Yale University Press.
- Kirby, Kris N., and Richard J. Herrnstein. 1995. "Preference reversals due to myopic discounting of delayed reward." *Psychological Science* 6 (2): 83–89.
- Manz, Wolfgang, and Helmut E. Lueck. 1968. "Influence of wearing glasses on personality ratings: Crosscultural validation of an old experiment." *Perceptual and Motor Skills* 27 (3): 704–704.
- Robbins, Trevor W., and Barry J. Everitt. 1999. "Drug addiction: bad habits add up." *Nature* 398 (6728): 567–570.
- Santilli, Paul C. 1983. "The informative and persuasive functions of advertising: A moral appraisal." *Journal of Business Ethics* 2 (1): 27–33.
- Scanlon, Thomas M. 1998. *What we owe to each other*. Cambridge: Harvard University Press.

- Spry, Amanda, Ravi Pappu, and T. Bettina Cornwell. 2011. "Celebrity endorsement, brand credibility and brand equity." *European Journal of Marketing* 45 (6): 882–909.
- "The science of shopping: the way the brain buys." *Economist* December 18, 2008.
Retrieved from: <http://www.economist.com/node/12792420>
- Thornton, George Russell. 1944. "The effect of wearing glasses upon judgments of personality traits of persons seen briefly." *Journal of Applied Psychology* 28 (3): 203.
- Williams, Bernard. 1981. "Internal and external reasons." In *Moral Luck*. Cambridge: Cambridge University Press, 101–13.

Journal of Cognition and Neuroethics

Philosophy and Neurobiology: towards a Hegelian Contribution on the Question of the Juridical Status of the Human Embryo

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, 2014. Volume 2, Issue 1.

Citation

Huesca Ramón, Fernando. 2014. "Philosophy and Neurobiology: towards a Hegelian Contribution on the Question of the Juridical Status of the Human Embryo." *Journal of Cognition and Neuroethics* 2 (1): 211–220.

Philosophy and Neurobiology: towards a Hegelian Contribution on the Question of the Juridical Status of the Human Embryo

Fernando Huesca Ramón

Abstract

Our purpose in this paper is to show the relevance of Hegelian Philosophy for discussion and reflection on Ethical and Bioethical matters, such as euthanasia, abortion, research on stem cells, genetic modification of human cells, etc. We shall deal, in the first place, with Hegel's notion of "person" (discussions on abortion or use of embryonic cells arise when one tries to attribute or deny *personality* to a determinate set of cells—that is when one tries to establish the juridical status of a developing human "embryo") as developed in his *Philosophy of Right*; then we shall deal with the contribution that Neuroscience can make to the understanding of the *material* or *natural* substrate, on which, ultimately, processes (mental, or *spiritual* in Hegelian language) related to the Hegelian notion of *person* rest. Finally, we shall offer a conclusion on the relation (first thinkable and experimentable in our own days) between the central nervous system and Philosophical concepts like *person* and *subjectivity*; in the end, it will be argued that Hegelian Philosophy offers an optimal model for an understanding of human freedom, will and rationality in terms of the neural activity of definite brain structures such as the limbic system, the prefrontal cortex, the basal ganglia, etc. Such Philosophical concepts, along with *personality* and *subjectivity* are essential when discussing and reflecting on the *personality* or *non-personality* of a human "embryo."

Keywords

Personality, subjectivity, Neuroscience, free will, freedom

Despise, if you will, understanding and science,
The highest of all the treasures of Men –
And to the Devil you will have surrendered
And must then perish.
—J.W. Goethe

A lot (and maybe not enough) has been written and discussed in the last year on subjects such as abortion, euthanasia, medical research on stem cells, cloning, utero-rent, fertilization *in vitro*, the uses of genetic information, etc. It has been said, on the one hand, that human life deserves an absolute and infinite respect; on the other hand, it has been said that the medical potential, in the sense of saving and improving the quality

of human lives is equally infinite, if it be the case that the juridical frame of a given territory allows the research potential of the biological sciences and techniques to be set free; likewise, on the public health arena there is the question of whether it should be allowed or not that a human being decides on the termination of his own biological life (euthanasia) or on the termination of the life of a human being developing in his (her—properly speaking) own interior (abortion); all this on the basis of a formalist (*à la* Kant) or utilitarian (*à la* Bentham) argument. Concerning all these discussions, one comes upon the following boundary concepts: “person,” “dignity,” “right,” “justice,” “legality,” etc. The quarrelling sides differ on the *conceptual* referent of such terms, and therefore, differ on the concrete *practical* agenda that must be defended and called for in the social and scientific spheres, on the basis of such conceptual referents. And finally, this quarrel has, as an arrival or resolution point, two concrete instances: the positive juridical reglaments and the collective frame of ideas (the habits and customs, ideology, etc.) of a given territory (a country, a province, etc.).

We consider that the present posfordist world calls pressingly for a renewed involvement of Philosophy with the medical, ethical, and legal debates of our days in such a way that, on the basis of the synthesis and systematization *pathos* that characterizes the philosophical thought of Plato, Aristotle, Kant, Adam Smith, Hegel, Marx, etc., it be possible to articulate the scientific breakthroughs in areas such as Molecular Biology, Genetics, and Neuroscience with philosophical subjects such as “personality,” “morality,” the “will,” “dignity,” etc.

So, our interest in this paper is to show the relevance of Hegelian Dialectics (dialectics, simply understood as a worldview which considers that from matter, there stems out consciousness, and from there the conceptual grasp which in turn allows the conformation of matter in accordance to ends) for reflection on and discussion of the subjects listed above concerning the treatment of the human embryo. In order to accomplish this, we shall focus on two central matters: first, the definition of the concept of “person” (because, in the end, when one reflects and discusses upon abortion or embryo-research, the dispute arises when personality is attributed or not to a given set of cells), and second, the contribution that Neuroscience can make to the understanding of the *material* or *natural* substrate which underlies, in the final count, the determinations around the concept of *person*. In the end, we shall offer a couple of concluding remarks on the relation (for the first time thinkable and experimentable in our own days) between the central nervous system and the philosophical concepts of *person* and *subjectivity*.

I

Hegel defines¹ the *person* as “the singular as free will”² (Hegel, 1983: 223). What is relevant in this definition, besides its *modern* character (in the sense of the History of Philosophy), is that it can be easily linked with current subjects of study and research in Neuroscience, which will be the subject of the next section. Therefore, we consider such a characterization to be the more adequate to the approach to the question on the juridical status of the embryo than the old and even canonical one from Boethius, which considers the person as an “individual substance of a rational nature” (Boethius *apud*. Agazzi, 2007: 115). In the final count, this definition, proceeding from a distant Medieval world, and of a theological and Christian inspiration (as well an Aristotelian), does not allow us to see clearly enough the decisive matter in the Hegelian perspective, on the question on *personality* itself, namely, the notion of *freedom*, and the concomitant notion of *will*.

In this moment, there arises the question: why should the Hegelian model of reflection centered upon “will” and “freedom” be preferred to the Boethian-Aristotelian one centered upon substantiality and rationality? To the purpose of our argument, we answer concisely: just the fact of stating this question (and as a matter of fact, any question!), is precisely already the affirmation of that, which, broadly, Hegel calls, *freedom*. In other words, in stating the question for the “being” of the substance, and for the ontological necessary conditions for the consideration of a given substance as “rational,” we are, above all, affirming a *theoretical* effort to *conceptually grasp* a given aspect of reality, or of the world in general, to guide, in a certain moment, our *practical* behavior. In this way, it should be confessed that behind every question (and to this day, we have only seen human beings, spontaneously, *state* questions) there is the affirmation of the freedom of *someone* who sets himself to question. To the concrete source of this *freedom*, Hegel calls precisely *will*.

Thus, we could preliminary conclude, in order to continue our reflection, that, in order to be able to state, firstly, any theoretical question such as the one of the “rationality” of a substance, one must have the *will* to do it. And the *will* to execute such a cognitive operation (of that later) cannot but be *free* (the end proposed—to grasp conceptually—cannot but be conceived by the inquirer himself). We consider that, in light of these brief

1. It may be remarked, that Hegel held lectures concerning the Philosophy of Right, from 1817 to 1831 in Heidelberg and Berlin. In this paper we use the *Philosophy of Right* from 1821 (the only one revised and authorized for publication by Hegel himself) and other secondary sources, based upon student notes from the lectures on *Philosophy of Right* held by Hegel.

2. All translations in this paper are ours.

statements, it should be possible to better understand Hegelian assertions such as “In this rests all the value of man: in that he knows himself as person” (Hegel, 1983: 44) and “The freedom of man, nevertheless, consists precisely in building his own nature, in making a nature for himself” (Hegel, 1983: 211). In this way, what is decisive and essential in the human being (as well as in his social, political and economical environment) is precisely, the *affirmation of freedom*, the construction of a *theoretical and practical* world in which he considers and knows himself as a free being, and acts as such (and so, not as a slave or as a serf).

The link of these reflections with legality (that is, with matters concerning “right”) is direct: “Will, in as much as it thinks itself [as free, one may add], is the source of right.” (Hegel, 1983: 209). So, one may conclude, that the source of right and so, of its main determinations as life, property and conviction is, fundamentally, free will, and not mere substantiality or natural materiality. Thus, following Hegel, it is not the natural constitution (and in our own days, we could say “genetic constitution”) which determines that the human being be a entity with absolute value and infinite; it is his spiritual constitution (that is, thinking, acting in accordance to ends, etc.) which does, because, indeed, only a thinking and acting being could declare itself and others as centers of absolute dignity and respect.

In this point we wish to introduce a transition to the biological part of our work, in order to, afterwards, take up the question on the juridical status of the embryo. The basis for establishing such a center of reflection is the following Hegelian fragment:

Entrails and organs are considered by Physiology as moments of the animal organism; nevertheless they constitute, in turn, a system of embodiment of the spiritual, and with that [in human being, above all], they attain an entirely different signification. (Hegel, 1991: 328)

Let us turn our attention to how Neuroscience can, presently, undertake this question, noted above of the “embodiment of the spiritual.”

II

“Free and conscious decision making, if at all existent, is one [of] the most complex presentations of human behavior. Process of decision making was frequently explored from the philosophical and psychological aspect, but remains [a] poorly studied topic in neuroscience” (Pirtosek, *et. al.*, 2009: 42) states a group of Slovene Neuroscientists in 2009. Also, in effect, the consideration of the subject of “free will” from the neuroscientific perspective is only possible towards the end of the twentieth century (Pirtosek *et. al.*,

2009: 38), and constitutes, therefore, an area of research, particularly recent and even in its origins.

What we intend now in this point is to (re)consider the old question of the relation soul (mind)–body, from a strictly immanent perspective, that is, a perspective that from the beginning rejects flatly the existence of two different substances (*res cogitans*, *res extensa* in Descartes), or of a “this side” and an “other side” (the Christian worldview and any other metaphysical ravings). Such a perspective is to be clearly found in Hegel: “The I determines itself [...] This is *freedom* of the will; freedom itself constitutes its own concept or substantiality; its center of gravitation” (Hegel, 1979: 55). If the I, or consciousness aware of itself³, or will related to itself, can determine itself, from itself—the reader will forgive the repetition—(and this itself, is placed within a *physiological* system of “the spiritual”), then, it is not necessary to invoke any instance such as *res cogitans* (Descartes), or an immortal soul (Christianity) to explain the translation of the will to determinate mental or motor *movements*. This is expressed by Neuroscience as follows: “From the non-dualistic perspective decision making is a brain process” (Pirtosek *et. al.*, 2009: 39).

So we can take forward our reflection to the following statement: “free will” (that which, as we saw above, determines from itself) has a physiological basis. So that emotional, motor and cognitive processes, which in the Hegelian system are studied under the titles of “inclinations and passions” (among others, as memory, etc.—see *Enzyklopädie §474*), “movement” caused by nerves (*Enzyklopädie §354*), and “theoretical spirit” (*Enzyklopädie §445*) have very concrete cerebral referents, as “mesencephalon,” “basal ganglia,” “lymbic system,” “prefrontal cortex,” which, in the final count, are responsible, precisely, of the transmission of “lymbic [that is, emotive], motor and cognitive information” (Pirtosek, *et. al.*, 2009: 42).

In this way, one may thoroughly speak of a “volitive system,” which integrates information “about a person’s specific needs and wants, personal and social norms for behaviour, current enviromental status, memories and effectiveness and consequences of past behaviour as well as a large body of additional information,” which may command concrete conducts such as “whether to act or not,” as well as the specific aspects of “what, when and how” (Drubach, *et. al.*, 2011: 243) of the action itself. In this way, this volitive system, even though it is not still completely characterized (Drubach *et. al.* speak to this

3. The Hegelian reflection of the “path” from consciousness to science can be found, in detail, in the *Phenomenology of spirit*. Hegelian “definitions” of *consciousness*, *self-consciousness*, *reason*, *spirit*, *will*, etc., can be found there.

moment of a “black box”—see Drubach, *et. al.*, 2011: 245), can indeed be associated to determinate areas of the cerebral cortex such as the dorsolateral prefrontal cortex, and the ACC (anterior cingulate cortex—see Drubach, *et. al.*, 2011: 245); the recent and concrete evidence for establishing a correlation between these areas and the “superior” functions associated with free will comes from specific clinical cases in which an alteration of the “normal” function of these areas (caused by mechanical damage or cell degeneration) produces effects such as “poor planning and/or judgement,” “alteration in “decision making,” “disinhibition, impulsivity, and altered goal-oriented action generation, implementation and retro-assessment” (Drubach, *et. al.*, 2011: 240). This has lead scientists, precisely, to consider certain areas of the brain (again, specifically the cortical region and concretely in the prefrontal cortex) as the “neurobiological basis” of “free will” (Drubach, *et. al.*, 2011: 239).

In this way, the initial Hegelian definition of person as “the singular as free will” should receive a new and intensified clarity, to the light of this neurobiological reflections. As, in effect, the thesis of the neurologists that “The voluntary action starts with determination of the purpose of the action” (Pirtosek, *et. al.*, 2009: 49) concurs perfectly with Hegelian theses that read “All the determinations of the will may be called ends” (Hegel, 1979: 55) and “A will which does not determinate itself is not a true will” (Hegel, 1979: 64).

III

It is moment to lay down some reflections, as a conclusion, in order, above all, to undertake finally, the question concerning the juridical status of the embryo.

Hegel and Neuroscience, coincide, broadly, in these two theses:

- The fundament of freedom lies on the *capacity* of the will to assume (that is, conceive, and execute or reject) determinate *ends*.
- The capacity of the will to assume determinate ends lies in the *physiology* of the (central) nervous system.

And so, if we follow the Hegelian characterization of the person as a “singular as free will,” we should already possess the sufficient elements to offer a couple of conceptual guidelines in order to consider which juridical status a certain cell group possesses, throughout its development *process*, as, in effect “The embryo is, *in itself*, a human being, it is not, however, *for itself*; for itself is the human being only as it is culturally educated reason which has *made* itself, what it is *in itself*” (Hegel, 1986: 25); so that the task of a Philosophy (or Ethics to be more precise) of Hegelian inspiration, combined with a neuroscientific approach, such as the one sketched above, consists, clearly and distinctly

in dealing with this question: in which specific moment of the embryonic development are the physiological structures to be found, *without which*, or without whose *adequate functioning*, the dialectical-neurological phenomenon of the *assumption of determinate ends*, does not take place?

To respond, even approximately, to this, one must consider the following: the “new” human being (embryo, in the general sense: in any case, the proper definition of embryo shall be given below) product of the fecundation, that is, of the union of a maternal ovule with a paternal spermatozoon, begins being only a mere aggregate of cells in constant differentiation and division. Only after concrete and determinate moments of development, the development of the physiological *origins* of the future organs and systems (fully developed) occurs, in a newborn baby, or in a child or adult. For instance, before the fifteenth of sixteenth development day, in the new human being one can not find the ectoderm, mesoderm and endoderm cell layers, which are the *original source* of all tissues, organs and systems of the fully developed human being; the formation of these layers is called *gastrulation*. Then, it is not until the period that comprises the third and eighth week of development when the process of *organogenesis* begins, that is, the beginning of the formation and differentiation of organs and systems. In this period, it may be properly spoken of *embryo*, while in the development staged comprised between the ninth week and birth, one should properly speak of *fetus*. Finally, structures such as the thalamus, the third ventricle, the mesencephalon, the brain stem and the cerebral hemispheres are not developed until the twelfth week of gestation (Sadler, 2003).

As one may tell, an important source of controversies arises precisely around the subject of the marked (or markable) out phases of embryonic development, and their relation with the notion of *personality* in the embryo. We, on the basis of the theory presented above, can contribute the following:

- Until before gastrulation, and organogenesis, the structural origins of the “system of the will,” do not even exist.
- It is until the twelfth week of development when one may speak of the structural existence of the “system of the will.”

In this way, we could conclude: an embryo, in the general sense of the term, before the third week of development does not possess the juridical status of a person in any way. Because it does not even possess *original* or *primitive* elements of the system of the will, necessary for the assumption, execution and rejection of ends; instances *without which* it may not be thought or spoken of *personality* in the proper sense. The question of the later embryonic development, without any doubt, may be yet a matter of controversy, as one would have to determinate, in the most possibly exact way, the moment of development,

in which the brain areas involved in the dialectical process of the will, are developed and articulated. The twelfth week date, to this date, seems to be widely accepted, as the moment in which the central nervous system shows a sufficient development, so as to be able to transmit emotive, motor and cognitive information; instances, it may be repeated again, necessary for the full expression of the will.

Finally, one may deduct the practical guidelines on the basis of these premises. For example, the right of women to interrupt pregnancy before the third week of development must be *tenaciously* defended, without any fear of damaging any right (as there is not even one neuron in the embryo,⁴ an element *without which* there does not exist a central nervous system *at all*, and so, neither *personality*, nor *subjectivity*, we may add). In the later stages one could endure, and even foster subsequent debates; however, one may, with some reservations, defend the right of women to interrupt pregnancy before the twelfth week of embryonic development (because, before this moment, the structural presence of the system of the will is not to be *fully* found).

Questions such as embryonic cell research, the extraction of stem cells from embryos, the *in vitro* cultivation of embryos for the purpose of research, etc., would deserve a special treatment; one that should include the Hegelian spheres of morality and ethicity. However, the juridical basis for reflection and discussion would be the same as the one expounded here.

4. See Sadler, 2003: 433.

References

- Agazzi, E. 2007. "El estatus ontológico y ético del embrión humano." In *Dilemas de bioética*. González Valenzuela, J., (ed.). Mexico: Fondo de cultura económica.
- Drubach, D., Rabinstein, A., Molano, J. 2001. "Free will, freedom of choice and Frontotemporal Lobar Degeneration." *Brain Mind and Consciousness: An International Interdisciplinary Perspective* 9 (1): 238–250.
- Hegel, G.W.F. 1979. *Grundlinien der Philosophie des Rechts, Werke 7*. Germany: Suhrkamp.
- Hegel, G.W.F. 1983. *Die Philosophie des Rechts. Die Mitschriften Wannemann (Heidelberg 1817/18) und Homeyer (Berlin 1818/1819)*. Germany: Klett-Cotta.
- Hegel, G.W.F. 1986. *Phänomenologie des Geistes, Werke 3*. Germany: Suhrkamp.
- Hegel, G.W.F. 1991. *Enzyklopädie der philosophischen Wissenschaften (1830)*. Germany: Felix Meiner Verlag.
- Pirtosek, Z., Georgie, D., Gregoric-Kramberger, M. 2009. "Decision making and the brain: Neurologists' view." *Interdisciplinary Description of Complex Systems* 7 (2): 38–53.
- Sadler, T. 2003. *Langman's Medical Embryology*. Philadelphia: Lippincot Williams & Wilkins.



cognethic.org