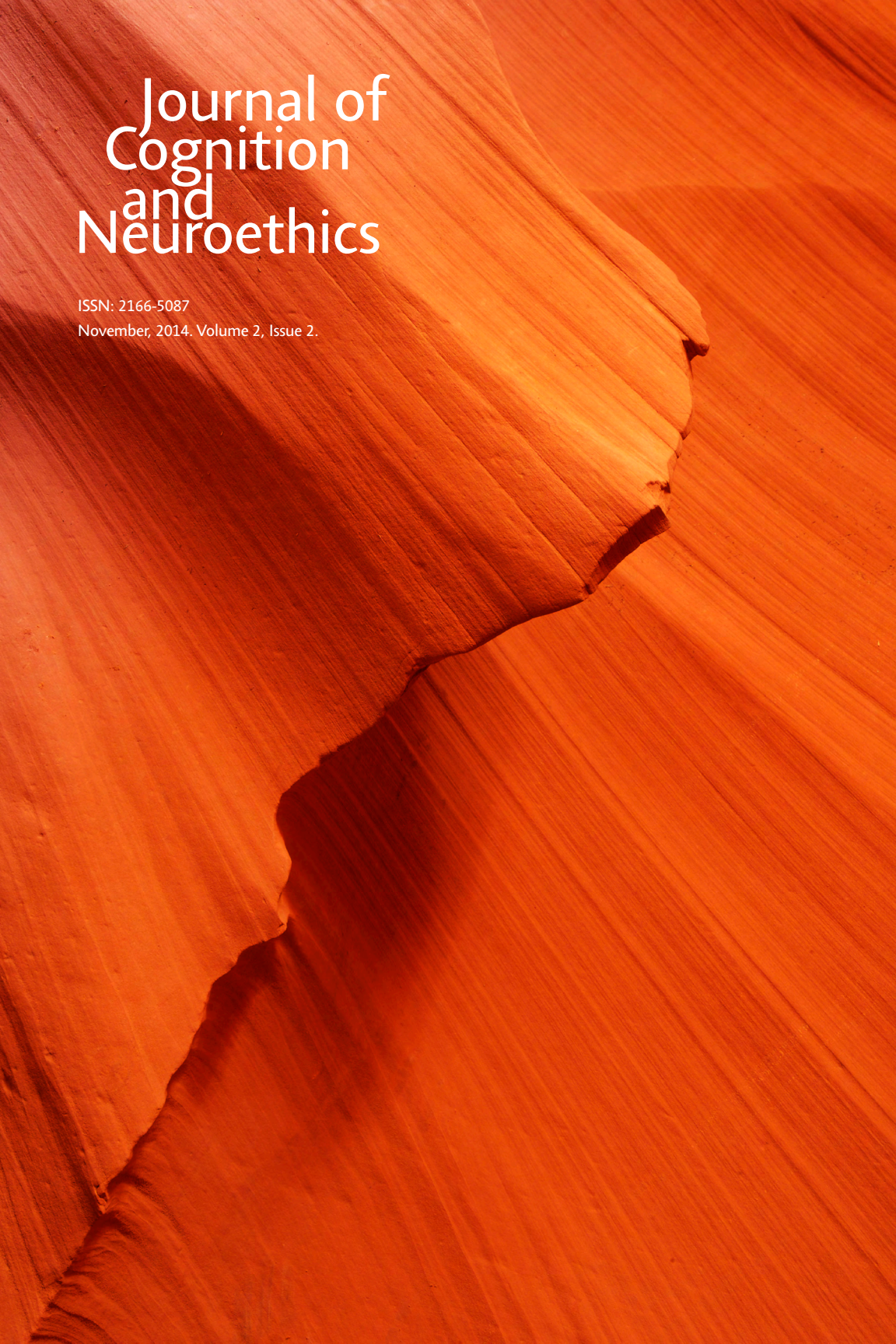
Journal of Cognition and Neuroethics

ISSN: 2166-5087
November, 2014. Volume 2, Issue 2.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 2, Issue 2 was digitally published in November of 2014 from Flint, Michigan, under ISSN 2166-5087.

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

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

Table of Contents

Journal of Cognition and Neuroethics

Cognitive Bias Modification as a Remedy for Weakness of Will

Brandon Gillette

Biography

Brandon S. Gillette earned an MA and PhD in Philosophy from the University of Kansas in 2013. Dr. Gillette is interested in investigating the ways that people process information and make decisions and value judgments. He currently teaches philosophy, in particular logic and ethics, at Maple Woods College, Johnson County Community College, and Washburn University.

Publication Details

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

Citation

Gillette, Brandon. 2014. "Cognitive Bias Modification as a Remedy for Weakness of Will." *Journal of Cognition and Neuroethics* 2 (2): 1–29.

Cognitive Bias Modification as a Remedy for Weakness of Will

Brandon Gillette

Abstract

There are two dominant perspectives from which to explain *akrasia* (weakness of will). *Akrasia* is intentional action against one's own better judgment, and paradigms that account for *akrasia* differ mainly in the role that normative judgments play in practical reason. The cognitive perspective regards *akrasia* as a cognitive defect, while the more common Humean perspective regards *akrasia* as an affective defect. In this paper, I argue that the cognitive account of akrasia is in better harmony with a variety of empirical findings in psychology and psychiatry. Further, I appeal to research on addiction, addiction recovery, and emotional disorders that indicates that *akrasia* is better remedied by treating it as a cognitive rather than as an affective phenomenon. Taken together, this provides a strong reason to prefer the cognitive account of *akrasia* to the more standard Humean model.

Keywords

Moral psychology, *akrasia*, weakness of will, cognitive bias, heuristic, addiction, cognitive bias modification treatment, normative judgment, irrationality, synchronic irrationality

There are two dominant perspectives from which to explain *akrasia* (weakness of will).[1] *Akrasia* is intentional action against one's own better judgment, and paradigms that account for *akrasia* differ mainly in the role that normative judgments play in practical reason. On one perspective, a normative judgment is a belief, and it is the sort of belief that normally influences action. Most of the time, when I believe that A is better than B, I choose A, and choose A *because* I judge that it is better than B. When I do not, it is because something has gone wrong with my beliefs or the way that I am processing information. This account of *akrasia* treats *akrasia* as a cognitive problem, so I will refer to this account as 'cognitive *akrasia*.'

The opposing (and more common) approach to explaining *akrasia* is to treat it as a desiderative problem. People often act as they judge best because most of the time peoples' desires are appropriately lined up with their normative judgments, either

1. The Greek term 'akrasia' has historically been translated into English as 'incontinence' or 'weakness of will.' The first carries with it unwanted associations, while the second seems passé, as few contemporary analytic philosophers talk of 'the will' in a traditional way. It has become standard to simply anglicize the Greek term, so that *akrasia* is the phenomenon in which an *akrates* behaves *akratically* or in an *akratic* way.

because their judgments are appropriately influenced by or else appropriately influence their desires. Because of the role that desires play as the sole bearers of motivation, this account is often called a Humean[2] account.

The chief difference between cognitive *akrasia* and Humean *akrasia* is that the cognitive view holds that normative judgments, regarded as cognitive states, cause motivation, while the Humean position regards motivation as essentially non-cognitive. In the case of *akrasia*, the truth of the cognitive account would imply that once the cognitive defect responsible for *akrasia* is modified or eliminated, *akratic* behavior is reformed, while truth of the Humean position would imply that only modification of desires would reform the *akrates*.

Suitably interpreted, this is an empirical question. In this paper, I have assembled evidence and explanation that shows that modification of cognitive states as opposed to conative or affective states is more reliably indicative of behavioral change in the *akrates*. Much of this evidence is taken from studies of addiction and clinical approaches to reforming addicts, so I shall begin this account by describing the relationship between addiction and *akrasia*.

Akrasia and Addiction

Akrasia brings about blameworthy actions distinct from some other related cases of blameworthy action. I wish to be specific about the blameworthiness of *akrasia* in order to discuss what it is that a person is expected to do in order to avoid the disapprobation justly due the *akrates*. To this end, I would like to bring in a distinction between intemperance, *akrasia*, and compulsion as they are differentiated by blameworthiness.

Intemperance: An agent pursues a course of action, *c*, that is objectively incorrect (i.e., that by some reasonable account he ought not to do), while making no normative judgment opposing his doing *c*. Intemperance is a failure to be motivated to behave as one ought. The agent makes an objectively poor choice, and is generally blamed for choosing so.

Akrasia: An agent pursues a course of action, *c*, that by her own judgment is not the best course of action open to her. The agent chooses irrationally and is generally blamed

2. David Hume, in his "A Treatise of Human Nature" and elsewhere developed the notion that conation (desire), as opposed to cognition (belief) had a primary role in explaining motivation. There are many current philosophical views inspired by Hume's basic stance. Here I use 'Humean' to group such views, though there are disagreements between proponents of these views. Further, there are contemporary Humean views that Hume himself may not have endorsed.

for choosing irrationally, but may be given some credit for knowing better and regretting *c*, unlike the unrepentant intemperate.

Compulsion: An agent cannot help but pursue *c*, whether judging that *c* is the superior or inferior option. The agent cannot alter his own compulsive behavior, and this is why it is called compulsion. The agent doesn't choose and is not generally blameworthy (except insofar as he allows the conditions for compulsion to obtain, and/or does/did not seek help in redressing his compulsive behavior).

This distinction is important chiefly because our approbative responses to each of these phenomena are different. In the case of the intemperate, we blame the lack of motivation to do what one ought. The way we go about reforming the intemperate is by convincing them of what they ought to do or not do, and if that fails, we generally try to motivate the intemperate through reward or punishment to assist them in ceasing to behave recklessly. The *akrates* is culpable for behaving as they themselves would condemn, but the fact that they themselves condemn it often gives partial credit. Instead of having to convince the *akrates* of the best way to act, we need only assist the *akrates* in attending to her better judgment. The compulsive is a case in which persuasion or another ordinary sort of motivational change is not effective. Generally, we regard compulsive behaviors as psychiatric pathologies of one kind or another, and attempt clinical interventions if the compulsion interferes with the subject's ability to live a reasonably satisfying life.

The cognitive account of *akrasia* distinguishes the three cases above on the status of their normative judgments vis a vis their actions. The intemperate simply has no normative judgment contrary to their action (but could have such a judgment and ought to). The *akrates* acts contrary to normative judgment (but could have and ought to have avoided such action). Compulsives cannot do other than they in fact do, whatever they judge, and so given the reasonable and common view that 'ought' implies 'can,' they are not blamed.

The Humean account explains *akrasia* in terms of the permanent irrelevance of normative judgments (understood as cognitive states) to motivation. Action against normative judgment occurs when someone judges that x is better than y but desires y more than x, and so does y. Under this view, the strongest desires supply motivational force, so people do whatever they most desire to do. Many have found this explanation of *akrasia* plausible (Mele; 1987; Stocker 1979). If this is the best explanation for *akrasia*, then reforming the *akrates* consists in some form of desire modification.

If the Humean explanation is really the best account of *akrasia*, then the approach that it suggests toward reforming the *akrates* should be the one that demonstrates the best success. If, on the other hand, a primarily cognitive approach is most effective,

it is reasonable to conclude that *akrasia* is a primarily cognitive problem. In other circumstances, this kind of reasoning bears fruit. If a mechanic replaces part A, and the problem is subtly affected, while replacing part B largely or completely fixes the problem, then the mechanic can reasonably conclude that part B is the largest part of the problem. What I intend to demonstrate in what follows is that success or failure in reforming frequent *akrasia* is actuated more by cognitive factors than desiderative factors.

There has been, to my knowledge (and I have searched extensively) no study of reforming anything called '*akrasia*,' so I am faced with the task of finding something that has been studied that matches the criteria for *akrasia*, even though the terminology under which it is studied in psychology, psychiatry, and physiology is not the same as philosophical terminology.

Akrasia per se has not been the subject of empirical study, but one sort of frequent *akratic* behavior that has received a great deal of empirical study is addiction. Instances of addiction as examples of *akrasia* are nothing new in philosophy, but even so, I shall go to some length drawing parallels between the cognitive account of *akrasia* and the cases of addicts continuing to engage in the addictive behaviors despite judging that they ought not.

The first step in this process is to recast the distinction between the intemperate, the *akratic*, and the compulsive (above) as a distinction between different sorts of addict. To this end, I appeal to a set of cases supplied by Gary Watson that instantiate the above definitions of intemperance, *akrasia*, and compulsion (1977, 324):[3]

1. The reckless or self-indulgent (intemperate) case: the woman who knows that having another drink will likely result in her becoming drunk and unable to fulfill other obligations, but who prefers the drink and accepts the consequences. She acts in accordance with her best judgment.
2. The weak (*akratic*) case: the woman who judges that it would be better not to drink, who could have refrained, but did not. She acts contrary to her judgment.
3. The compulsive case: the woman who judges that it would be better not to drink but who was unable to refrain. She also acts contrary to her judgment.

It squares with common experience that addicts are often, at various times, one of the three above. Of course, for my purposes, I shall focus attention on the addict who judges that they ought not behave as they do, and who is capable of avoiding that behavior.

3. I do not know if Watson, if pushed, would accept Humean normative judgment externalism, but his account in this piece does seem to take seriously the main theoretical commitments of the Humean perspective. See (Smith 2003) for commentary on Watson's distinction.

The example Watson brings in has to do with choosing to drink another drink of alcohol. This is appropriate, as many instances of failing to do as we judge that we ought are often bound up in (at the mild end of the spectrum) bad habits or (at the more severe end) very serious addictions. Failures to change our habits, like when starting a new diet, are frequently cited as candidates for *akrasia*. In the context of discussing addiction, I shall provide an account of cognitive bias modification along with evidence of its effectiveness in assisting persons in breaking addictions. Again, this is an important part of the account because it demonstrates that the sort of weakness involved in *akrasia* is a cognitive weakness because it is most effectively remediated by addressing its cognitive aspects. In identifying the correctable weakness implicated in at least these cases of *akrasia*, I shall be identifying the weakness that the *akrates* is culpable for failing to correct.

Before I begin with the main discussion, I would like to point out an issue that may arise that might make the following account more likely to be misunderstood. I have made extensive use of empirical data from psychology in developing a cognitive account of *akrasia*, and I shall make use of literature primarily from psychiatry in detailing cognitive bias modification as it pertains to reforming *akratic* behavior. I do not thereby mean to give the impression that I regard *akrasia* to be a pathology requiring clinical intervention. I hold what I believe is a common view of pathologies requiring clinical intervention. That is, I view such pathologies more like examples of compulsion rather than *akrasia* or (at least ordinary) intemperance.

In fact, in denying the existence of *akrasia*, some have characterized reported cases of *akrasia* as instead being cases in which the dictates of a person's best judgment are psychologically impossible for her to follow (Hare 1963).[4] In any of these cases (some of which surely must exist) the action that takes place against better judgment is not intentional, and thus is not *akrasia*. It is instead a version of psychological compulsion. In adapting psychological literature to empirically inform the philosophical concept of *akrasia*, it would be tempting to identify *akrasia* with an existing mental disorder.[5] I encourage the reader to resist such temptation, and regard clinical pathologies as more akin to instances of compulsion (in the sense that compulsion operates in Watson's example) than to *akrasia*. Even if psychologists and psychiatrists do not generally regard pathological behavior to be unfree, the more common view of praiseworthy or blameworthy action involves action that is suitably under the control of the individual

4. See especially Hare's Ch. 5. See also (Hardcastle 2003) for a critique of an attempt to reduce psychological explanation of *akrasia* to neuroscience.

5. This is largely what (Kalis, et al. 2008) attempt.

in question.[6] I shall avoid entry into any debates concerning metaphysical free will. The common view may or not be ultimately mistaken about pathological behavior, but at the very least I shall be able to provide an account of cognitive bias modification that addresses *akrasia* but that is not a strategy that requires or is restricted to clinical intervention.

An addiction is a pattern that persists over time, while *akrasia* is episodic. Of course some are *akratic* more often than others, and addictive behaviors will be correlated with more frequent occurrences of *akrasia*. However, not all addictions are created equal. There appear to be many different sorts of addictions, some involving chemical dependence, and the strongest of these may appear better examples of compulsion than *akrasia*. Also, some addictions, like my own utter dependence on my morning coffee, are, if anything, examples of intemperance rather than of *akrasia* as most people do not care to break their mild to moderate caffeine addictions.

I shall like to leave aside both these most severe cases of addiction and the mild addictions that people generally don't regard as particularly bad or worthy of effort in breaking. I contend that there are sufficiently many examples of addicts who are capable of controlling and/or breaking their addictions, judge that it would be best to do so, and still sometimes fail to perform the individual actions that eventually lead to the breaking of a bad habit. These phenomena are rather well studied.[7]

Common experience tells us that at least some addictions that addicts wish to break involve instances of action against better judgment. It is part of our common knowledge of alcoholism, for instance, that most admitted alcoholics do not think it best that they continue to be alcoholics.[8] As a necessary step in demonstrating that cognitive bias modification is an effective remedy for *akrasia*, I must demonstrate the role that cognitive bias plays in those addictions that include examples of *akrasia*.

Cognitive Bias in Addiction

Two kinds of cognitive defects are at play in both substance addictions and behavioral addictions that do not involve substances. One of these defects is a cognitive bias that

6. Aristotle, NE book III, agrees.

7. See (Campbell 2003), which is one of a very few articles that specifically identifies addiction with "*akrasia* or weakness of will."

8. For voluminous examples of this see: (Alcoholics Anonymous 2001), or any other collection of testimonies of alcoholics or recovering alcoholics.

minimizes recall of the negative effects of the addictive behavior. Let us refer to this as recall bias.

Typically, rewards for behaviors tend to reinforce those behaviors, while negative consequences for behaviors tend to discourage repetition of those behaviors. Long experience with conditioning, incentives, and disincentives tells us that such a connection is as regular and reliable as any psychological law. Objectively, addictive behaviors are often harmful. A failure of the addict to reform his or her own behavior, or even to recognize the problem, is a cognitive failure—a form of subjective irrationality as well as a failure of objective rationality.[9]

Akrasia is a failure of synchronic rationality, and not a failure of diachronic rationality.[10] If addicts believed that their addictions were not harmful or if they misestimated the consequences of their addictive behaviors in a way that additional information or a different way of considering things would fix, then the addict would demonstrate a failure of diachronic rationality. Surely this is what happens some of the time, but it does not capture the full range of mental processes often associated with the persistence of addictive behaviors. Those who seek to give up their addictions often do so on the basis of the past negative consequences of addictive behaviors. It is the failure of their own past negative experiences to sufficiently motivate addicts that is, in a way, paradoxical. Being able but not disposed to remember negative consequences of addictive behaviors fits well with Aristotle's talk of having but not attending to knowledge,[11] as well as the more empirically respectable talk of information that is or is not present in the global workspace (Baars 2003).

The phenomenon of recall bias makes a charge of subjective synchronic irrationality (*akrasia*) intelligible and empirically verifiable. It is not that the addict believes things about their addiction that are false, or that they must revise. Instead, the past negative consequences of the addictive behavior are often not recalled at all.

---

9. For more on the distinction between subjective and objective rationality see (Wedgwood 2003).

10. Synchronic irrationality is marked by consistency of judgments and intentions as recognized by the subject. Diachronic irrationality is a failure in the rules and procedures for forming judgments. For more on this distinction, see (Wedgwood 2007).

11. Aristotle is the first thinker in Western Philosophy to take seriously the notion that people may sometimes in fact pursue what they themselves acknowledge as a lesser good. Aristotle's account of *akrasia* is notoriously difficult, but tends to focus on how someone might have knowledge of the good and yet fail to attend to it. See Book VII of his *Nicomachean Ethics*.

William Campbell, a fellow of the American Society of Addiction Medicine, is one of a few who explicitly link *akrasia* in the case of addiction to a specifically cognitive impairment (recall bias). Campbell describes the causal relevance of cognition in addictive behavior as follows (the italics are my own):

> Addicts appear to be acting at various times on 2 different belief systems. The first belief is that the addictive behavior is harmful and produces negative consequences…The addict appears to act on the basis of faulty reasoning, and the actions are such that *cognition does not appear to consider* the previous negative consequences of the addiction. (671)

This is a clear description of *akrasia* and its classification as a failure of subjective synchronic rationality. At this point it is tempting to ask what feature of addiction causes this lack of recall. This is a subtle confusion. It is like asking what feature of forests is responsible for causing trees to clump closely together. Campbell is arguing for cognitive bias as a causally necessary aspect of the etiology of addiction. It is not just clinicians who appear to hold this view. Campbell cites some literature from Alcoholics Anonymous, an organization with a wealth of practical experience that should not be discounted. In particular, Campbell singles out the statement that "…we shall describe some of the mental states that precede a relapse into drinking, for obviously this is the crux of the problem" (671).

Recall bias is not the only kind of cognitive or attentional bias implicated in addictive behavior. Another sort of cognitive/attentional bias implicated in addiction occurs in the increased attention to addiction-related stimuli in the addict. We may refer to this as focus bias. Focus bias and recall bias serve together to make the addict more aware of the presence of temptation and less cognizant of its previous bad consequences. All human beings make implicit use of heuristics (short-cuts that make decisions on less than the total amount of available information) when making decisions (Cosmides and Tooby 1994; Gigerenzer and Goldstein 1996; Gigerenzer and Todd 1999). These heuristics are generally useful to us, but in some contexts can be misapplied and supply the wrong decision. The heuristic gets labeled a 'bias' when it gets misused. Consider the ordinarily useful traits of selective attention and memory. Having our attention drawn to the fastest moving object in our surroundings can have survival value. Often, fast-moving things are dangerous (charging predators) or else are opportunities for food (fleeing prey). Calorically dense food items present themselves readily to the attention because there has historically been value in knowing where the calories in our environment are. Generally, the ability to

see what we want more readily than what we don't want is very useful. In the context of addiction, such tendencies are positively and powerfully counterproductive. It makes sense on these lights to regard addiction as a misapplication of the ordinarily useful cognitive tools that are selective attention and memory.

Medical and psychological researchers, in studying addicts and their characteristic behaviors, have noticed a number of ways in which addicts of various kinds share cognitive traits. These traits have become an integral part of understanding the cognitive aspects of addiction. Focus bias, as it is studied, consists in the following: a tendency of addicts to respond to certain cognitive cues more quickly than non-addicts, a reduced tendency of addicts to disengage attention from addiction-related cues and onto non addiction-related cues, and a reduced tendency compared with non-addicts to distinguish target cues from distracters (Mazas, Finn and Steinmetz 2000).

It is significant to recognize that these same sorts of cognitive biases contribute to a startlingly wide range of addictive behaviors, which includes both addictive behaviors that do and addictive behaviors that do not involve any psychoactive or mood-altering substances.[12]

A wealth of evidence suggests that increased attentional bias toward addiction-related stimuli predicts relapse of addiction among a startling diversity of addictions. As one example of attentional bias in addicts, a study by Liu et al. made use of what is known as a Stroop task to demonstrate focus bias in cocaine addicts (Liu, et al. 2011). In a Stroop task, cocaine addicts and controls are contrasted in their abilities to identify the color of a word while ignoring the word's meaning. The word is presented, and subjects (both cocaine addicts and controls) are asked to press color-coded buttons corresponding to one of the potential colors of the presented words as quickly as they can accurately do so. Some of the words are cocaine-related (e.g., 'cocaine,' 'dealer,' or 'freebase') while an equal number of words are neutral with regard to cocaine and length-matched with the cocaine-related words (e.g., 'cabinet,' 'window,' and 'armchair'). A significant difference in cocaine addicts' reaction times to neutral versus cocaine-related words is evidence of attentional bias to cocaine-related stimuli. Controls show no significant difference in reaction time to cocaine-related versus neutral stimuli. The Liu et al. study confirmed the results of other studies (Hester, Dixon and Garavan 2006; Vadhan, et al. 2007) that find an increase in what is above termed 'focus bias' among cocaine addicts.

12. Though Seamus Decker and Jessica Gay claim that research into the role of cognitive bias in addiction is scarcer for "evidence about behaviors that do not involve drug use or other physiological factors." See their (Cognitive-bias toward gaming-related words and disinhibition in World of Warcraft gamers 2011).

Further, Liu et al. write "[I]mproving impulse control and remediating attentional bias may prove to be helpful tools in the treatment of cocaine dependence" (2011, 121). It stands to reason that if the remediation of cognitive bias would assist in the treatment of cocaine dependence, other sorts of chemical addictions should admit similar amenability to cognitive bias modification as effective treatment. Some evidence confirms this suggestion, indicating that higher attentional bias negatively correlates with the success of treatment outcomes for alcoholics (Cox, et al. 2002) and similar confirmation in the case of smokers (Janes, et al. 2010). The Janes et al. study is particularly interesting. The study measured brain reactivity to cues related to cigarettes and to smoking, and concluded that there was a strong negative relationship between brain reactivity to smoking-related cues and likelihood of continued tobacco abstinence among smokers who wish to quit smoking (our *akratic* addicts). They also found a correlation between brain reactivity (measured by fMRI data) and attentional bias (measured by a Stroop task). In concluding "…that prequit brain reactivity to smoking-related images is greater in smokers who eventually slip after attaining brief abstinence with NRT and that anterior insula and dACC fMRI cue reactivity correlate with an attentional bias to smoking-related words." Janes et al. provide a neurological confirmation of the role played by attentional bias in addictions.

The empirical evidence for the important role that attentional bias plays in addictive behavior is not restricted to chemical addictions like alcoholism or addictions to cocaine or tobacco. Other studies have uncovered similar attentional bias (characterized by focus bias and recall bias) among overeaters (Nijs, et al. 2010), pathological gamblers (Boyer and Dickerson 2003), and computer gaming addicts (Decker and Gay 2011).

Decker and Gay, studying computer gaming addiction, used an Affective Shifting Go/No-go Task (ASGNG) to measure cognitive bias toward gaming-related cues among habitual players of a particular video game against a control group of non-players. The ASGNG (not an abbreviation that rolls off the tongue easily) task is similar to the Stroop task. A set of positively valenced common English terms as well as positively valenced jargon specific to the video game are targets, while negatively valenced English and jargon counterparts are distractors for some trials, vice-versa for other trials. Subjects are asked to identify the targets by pressing a button when they are displayed, and are instructed not to press the button for the distractors.

So each subject would be expected to press the button for a word like 'friend,' a positively valenced English word, as well as for 'purple,' a positively valenced word for

World of Warcraft players.[13] Subjects would likewise be expected to leave off the button for negatively valenced English or World of Warcraft phrases, like 'betray' or 'nerf'[14] respectively.

The World of Warcraft players demonstrated cognitive bias toward game-related stimuli by more quickly and accurately distinguishing between game-related targets and distractors than English targets and distractors, and also distinguished game-related targets from game-related distractors more quickly and accurately than the control group of non-players distinguished English targets from English distractors. Decker and Gay conclude: "Similar to past research showing that recovering alcoholics had cognitive-bias to alcohol-related words, [game players] with high rates of time spent playing computer games showed cognitive-bias toward gaming-related words" (807–808).

It has long been clear that cognitive performance can be habituated—practice enough memorization and you will become better at memorizing things, even without intentionally trying to do so. The role of habit and cognitive bias in the case of the addict seems to be a kind of feedback loop. The addict trains herself to recognize and seek addiction-related stimuli, and this makes the attentional bias toward addiction related stimuli stronger. If attentional bias really is as central to addiction as the evidence suggests, this feedback loop would explain why those who have been addicted for a greater period of time find it harder to break an addiction. The attentional bias is more highly habituated in the long-term addict.

Because the same forms of cognitive bias are observed accompanying so many varieties of addiction, it is reasonable to postulate that these cognitive biases are central to what we mean by 'addiction.' Evidence that the degree of cognitive bias varies concomitantly with the strength of the addiction (measured in rates of abstinence from the addictive substance or behavior) is further reason to believe that cognitive bias is an essential element of addiction. Since addictive behavior is often contrary to the better judgment of the addict, addiction provides a rich field of examples for the cognitive bias account of *akrasia*.

13. The most powerful and desirable pieces of weaponry and armor in World of Warcraft are most easily distinguished by their names written in purple text (for rare or epic items) versus blue (for merely uncommon items) or green (for run-of-the-mill items). Players refer to receiving such an item as, e.g., 'getting a purple.'

14. Blizzard, the company that maintains World of Warcraft, often makes changes in the abilities of certain classes of players' characters. Such changes that serve to make a class of character relatively more powerful are known as 'buffs' while such changes that make a class less powerful are known as 'nerfs.'

It is worth noting that in the philosophical tradition, examples of people wanting to change their behavior but failing to do so often involve bad habits or addictions. Unifying an empirically informed account of *akrasia* with empirical evidence concerning the role of cognition in sustaining addictions is a philosophically and scientifically significant development. It is philosophically significant because it is the first appearance of a thoroughly empirical account of a *long*-discussed phenomenon. It is scientifically important because it serves to unify separate avenues of research under a broader aegis. Given a clear empirically informed account of *akrasia*, the interested empirical researcher has a starting point in further studying *akrasia* as such, rather than inadvertently revealing elements of *akrasia* while studying addictions, cognitive biases, or decisional heuristics.

As I am primarily interested in the philosophical importance of an empirically informed account of *akrasia*, I shall briefly point out how the empirically informed account contributes to, and in some sense completes prior philosophical perspectives on *akrasia*.

In Aristotle's diagnosis of *akrasia*, undertaken to refute the position that *akrasia* is psychologically impossible, he proposes that *akrasia* is the result of having but not attending to knowledge of the good. Lacking the vocabulary of modern behavioral psychology, Aristotle appears to have anticipated, albeit in a very general way, the empirically informed explanation of *akrasia*. Replacing vague notions of having but not attending to knowledge with detailed empirical accounts of cognitive/attentional bias preserves the spirit of Aristotle's feeling concerning an appropriate explanation for *akrasia* and adds an empirically verifiable phenomenon on which to ground an explanation of *akrasia*.

Similarly, Donald Davidson, in developing an account of the logical possibility of *akrasia*, relies on a distinction between all-out judgments (judgments that consider everything relevant to the evaluation) and judgments with a prima facie operator that take the form pf(x is better than y, r) where r is the evidence considered.[15] Davidson does not consider (as it is outside the scope of his paper's limited purpose) whether the difference between all-out judgments and prima facie judgments is empirically verifiable. The empirically informed account of *akrasia* that I have been advocating fills this gap in this overall story of *akrasia* as well as Aristotle's. Because human beings are incapable of simultaneously considering all relevant evidence at the same time, and frequently act upon judgments of the form outlined above, it is clear that we ought to see cases of action based on evidence that is more apparent or that is attended to *first* (prima facie

15. While *akrasia* has never been a dead issue in philosophy, much contemporary discussion of *akrasia* has been inspired and influenced by Donald Davidson's landmark paper "How is Weakness of the Will Possible?"

judgments) than judgments based on evidence that a more patient thought process would reveal as superior.[16]

The idea that a specific cognitive *weakness* explains the difference between the addict who sincerely judges that they ought to break their addiction and still relapses has both commonsense currency and also empirical verification. If a computer gaming addict (for example) is more apt than the non-addict to take notice of gaming-related stimuli, and also apt to respond more quickly to gaming related stimuli than the non-addict, then it should not be surprising that their decisions concerning computer gaming are more frequently made on the basis of prima facie judgments with gaming-related stimuli crowding out non-gaming-related stimuli, accompanied by a failure to recall past negative consequences of excessive gaming.

Treating Akratic Addiction

William Campbell, mentioned above, approaches the problem of addiction and *akrasia* from a treatment perspective. Campbell is motivated by what he sees as a problematic lack of a unifying definition of addiction that explains why chemical addictions (like alcohol and cocaine) should have so much in common with behavioral addictions (like gambling).[17] Campbell argues that the field of addiction treatment has been held back both by lack of a comprehensive etiology of addiction, and by an "accepted view" that treats addiction as primarily conative. He puts it briefly: "The accepted view is that craving causes the addict to act" (671). Campbell follows this claim with a brief refutation of the conative accepted view. First, if the craving were causative, then every time the cravings became sufficiently strong, an abstinent addict would relapse. In reality, sometimes they do, sometimes they don't. Further, sometimes addicts who experience severe craving stop their addictions. These events tell against cravings as a sufficient condition for relapse or as a necessary obstacle to recovery.

This is not an extended argument, and it is a bit simplistic, but I think Campbell's point has merit, particularly since the previously discussed evidence indicates a much more central role for cognitive states in addiction than conative states. But because the conative view is so prevalent, it is worth more detailed examination.

16. See (Davidson 2001, 40) for a formal description of better reasons supplanting inferior ones in judgment.

17. "The present conceptualization of addiction inadequately explains addiction as an entity unto itself and does not provide any understanding of the relation between the substance and behavioral addictions" (Campbell 2003, 671).

One might preserve the conative view against Campbell's argument and posit that whenever the desire to quit is strong enough, it can overpower even the strongest of cravings, and when it is weak enough, it can be overcome even by mild cravings. However, this idea (though common) has its own conceptual problems. The will (in this case, whatever accounts for the desire not to be an addict) is often taken to be the feature of psychology that resists or fails to resist desire, and the 'will versus desire' description of *akrasia* has been historically prominent enough to translate *akrasia* as "weakness of will."[18] Watson, who is skeptical of the view, puts the problem this way:

> This talk of strength of desires is obscure enough, but insofar as it has meaning, there does not appear to be any way of judging the strength of desires except as they result in action…Isn't the only relatively clear measure of strength of desires [versus strength of the will] the tendency of those desires to express themselves independently of the agent's will?...If a sufficient condition of compulsive motivation is that the motivation be contrary to the agent's practical judgment, then weakness of will is a species of compulsion. (1977, 327–328)

In other words, the "will versus desire" picture of *akrasia* has difficulty distinguishing *akrasia* from compulsion. If some desire is so strong that nobody could overcome it, then it is a clear case of compulsion, but the evidence for this circumstance is identical to the evidence for someone with an extraordinarily weak will succumbing to a stronger, but still very weak (that is, resistable) desire.

Aside from this issue, the "will versus desire" theorist is constrained by their view to offer one of two remedies for the *akratic* addict. That is, the "will versus desire" theorist must provide some account of what it means to intentionally strengthen one's own will or else to intentionally weaken one's desires (both of which are themselves "acts of will"). I need not belabor the inherent circularity of using one's will to strengthen one's will. Put into layman's terms, the addict who judges that they ought to quit and is unsuccessful in quitting needs to find a way to either want the addiction stimulus less or else to want to quit more. Such a view is dependent upon some successful method of desire modification.

---

18. See also Davidson, "How is Weakness of the Will Possible?" p.27. Here, Davidson also characterizes a separation of "thinking we ought" and "wanting to" as the most common way of handling *akrasia*. This can legitimately be called the "received view" of *akrasia*. In (Paradoxes of Irrationality 2004, 175) Davidson expresses a similar worry to mine that the "will versus desire" picture (he calls it the Medea Principle) does not adequately distinguish *akrasia* from compulsion.

Despite the status of the "will versus desire" view as the received view of *akrasia*, it appears that few actually endorse the view in its entirety, while many argue against it. I have no intention of building up the naïve "will versus desire" view[19] because my primary opponent is the Humean. I bring up the "will versus desire" view because it shares one particular problem with the Humean, and that is how to best account for reform of the *akrates*.

If the *akrates* is to act in accord with their judgment that x is better than y, then the Humean must come up with an account for desiring x more strongly or desiring y less strongly. This is what Campbell has in mind in referring to a conative approach. It is my intent to show through additional evidence and analysis that it is much more productive to approach addiction from a cognitive angle than from a conative angle, and that the cognitive approach to reforming *akrasia* has been attended with greater success than the conative approach. My interpretation of the evidence is that normative judgments, understood as cognitive states, have a much greater role in normative motivation than the Humean can accept, and so my version of cognitive *akrasia* is the correct view.

Changing behavior without changing desires

Odysseus wished to hear the sirens sing, because their singing was said to be so beautiful that men would dash their ships upon the rocks pursuing the sirens who sang so. Knowing that his desire to pursue the sirens would be irresistible, Odysseus ordered his sailors to tie him to the mast and then to seal their own ears with wax, and not to let him loose until they were well clear of the sirens. As the story goes, Odysseus begged and pleaded and shouted for his men to untie him or to remove the wax from their ears, but they did not hear him, and followed his original orders. So Odysseus changed what would have been his behavior without changing the desire to pursue the sirens. He did this by recognizing his interests, anticipating his future desiderative states, and then manipulating his environment to make the pursuit of an irresistible desire impossible so as to act in accord with his better judgment.

Examples of this combination of foresight, careful judgment, and manipulation of our future selves can be termed 'Odyssean self-control' in his honor.[20] People take similar, though less heroic measures every day. Not keeping candy bars in the house so as to

19. See the latter half of (Watson 1977) for an attempt at this.

20. I first encountered this phrase in (Pinker 2011, Chapter 9).

avoid overindulging, not shopping for food while hungry, or seeking out a less distracting environment in which to work are all examples of Odyssean self-control.

Consider a more contemporary example germane to the current discussion. Ingrid is a recovering alcoholic. Let us stipulate that she is an addict who judges that it would be best not to be an addict, and so is an *akratic* rather than intemperate addict if she resumes drinking. She very much desires to drink, but of course has no trouble refraining from drinking while at work, as there is no alcohol available. Similarly, her husband helps her to ensure that she resists the temptation to keep any alcohol at home. The most direct route home from her workplace takes Ingrid by a pub where she has spent many an after-work hour drinking and socializing with her friends, some of whom she has had to break contact with because they have been insensitive to her efforts to stop drinking. She has even had her husband replace the phone numbers for these friends with the number of the local AA support line in her phone. Because she finds the temptation to stop at the pub nearly irresistible, she has stopped driving herself home from work, going as far as to sell her car and allow her driver's license to expire, replacing it with a mere government ID card. She takes the bus home, and there is no bus stop near her old pub.

The reason she goes to such heroic measures is to ensure that she would have to go to equally heroic measures to have a drink. She would have to solicit someone's cooperation, which might not be forthcoming if they know she is a recovering alcoholic, and she tells everyone she knows that this is the case. She would have to call and schedule a cab or walk a long distance to get to her old pub, and both of those are actions that give her much time to reconsider or not follow through with these plans in the course of her ordinary workday. In other words, these obstacles to drinking and going to the pub allow Ingrid sufficient opportunity to attend to her meta-judgment as opposed to being in a situation in which recall bias and focus bias would have a significant causal role in her behavior. This is a good example of Odyssean self-control, and what is most notable is that it is an attempt to modify behavior not by diminishing the desire to drink, but by Ingrid's reasoning out her likely response to environmental cues and then placing barriers in the way of encountering the cues likely to contribute to a relapse, while replacing some cues with cues likely to contribute to abstinence.

It would beg the question to say that either *of course* there is some desire not to drink in operation the whole time or that *of course* her judgment that it is best not to drink is sufficient motivation for her to work out the strategy that she has worked out. I do not deny that the Humean may be correct and that desires may hold a monopoly on motivation, but I think that Ingrid's case is one which, if taken at face value, demonstrates a form of cognitive self-manipulation. Whether there is some desire at play that counters

the desire to drink or not, Ingrid's strategy is essentially one that is actuated on her ability to anticipate consequences and manipulate her surroundings to achieve results that she judges best. These are cognitive abilities. Further, her efforts are all steps that are intuitively consistent with her judgment that it is better to quit drinking, while a failure to do something to keep herself away from bars and alcohol would be intuitively inconsistent with her better judgment, opening Ingrid to the charge of subjective synchronic irrationality.

Consider only one more fabricated example. Alex judges that it would be best to quit wasting so much time playing video games. He decides to make use of the best behavior modification research available and visits the website www.stikk.com.[21] The Stikk system was born out of credible research on incentives and behavior modification, and chiefly makes use of the insight that it is more effective to give someone a reward (say, money) and then threaten to take it away if the subject doesn't complete a goal than to offer the same reward only once the goal is completed. Alex, in order to make the Stikk contract, must set his goal: no more than ten hours of video gaming per week (hey, it's a start). Alex must then supply stakes. Most people choose to put money on the line, but the site allows a commitment contract without monetary stakes. Alex designates $10 for every hour exceeding 10 per week of video games that he plays. Of course, those at Stikk do not wish to profit off of others' *akrasia*, so the disincentive for failure has an interesting twist. If Alex's credit card must be charged, the money goes to the Westboro Baptist Church, whose views and practices Alex absolutely detests (the subject chooses their own anti-cause). Alex then selects a referee, who keeps track of his progress. Alex's roommate, Beavo, who has been most vocal about the amount of time Alex has been wasting at video games, is the logical choice. Finally, Alex enlists several friends and family members to act in the role of supporters, whom he keeps informed of his progress and from whom he receives encouraging feedback.

If this method of behavior modification works, as the laboratory work on which the method is based would suggest,[22] then it is also an example of a form of self-manipulation that relies on the ability of the individual to predict their responses (including what their desires *will be*) in counterfactual scenarios. Would it be most accurate to say that Alex stopped playing so many video games because he hated the Westboro Baptist

21. This is an actual website, founded by Ian Ayres, Dean Karlan, and Jordan Goldberg, two Yale economics professors and a former Yale student, respectively.

22. The site, as of June 3, 2014, lists just over 300,000 workouts completed and over 2.5 million cigarettes not smoked.

Church more than he loved video games? That makes some degree of sense, except that presumably Alex always hated the Westboro Baptist Church more than he loved video games, and that his hatred only mattered after he intentionally set up a system in which one was set directly opposed to the other. The cognitive anticipation of his future states is doing a great deal of motivational work. Even if the Humean is correct and desires hold a monopoly on motivation, there is at least room to pay much greater attention to cognitive states in a credible story of motivation, especially a person's normative judgments.

As far back as Aristotle, the difference between the *akratic* and the intemperate is couched in their actions relative to their best judgment. The intemperate chooses in accord with their own best judgment (and thus are subjectively rational), but are disapproved of because their judgment runs afoul of some objective standard (and thus are called objectively irrational). This distinction has some consequences that are relevant here. The intemperate person might be persuaded to *change* their judgment, or might not, but the *akratic* is susceptible to correction of their behavior by simply having their best judgment more readily brought to their attention.[23]

Even in cases in which we do little or nothing to change desires that we have, we may change behavior. In the psychological literature, cases like the above are termed 'cognitive bias modification'. The term sounds more clinical and impressive than it really is. Actually, the kinds of strategies employed in the various forms of cognitive bias modification in the literature strongly resemble the above two examples of Alex and Ingrid.

Cognitive bias modification treatments have their genesis in research aimed at treating various sorts of anxiety and depressive disorders. A significant part of the etiology of anxiety and depressive disorders involve certain cognitive biases, and indeed these biases are common across many emotional disorders. As Matthews and MacLeod put it in their literature review:

> Evidence has continued to show that, relative to emotionally stable individuals, those prone to emotional disorders preferentially attend to emotionally congruent cues, recall more unpleasant memories, and interpret ambiguous events in a more negative manner. The findings we have reviewed suggest that these emotional processing biases occur across emotional disorders, as perhaps might be expected in view of

23. Aristotle puts it "Moreover, the incontinent person is the sort to pursue excessive bodily pleasures against correct reason, but not because he is persuaded [it is best]. The intemperate person, however, is persuaded, because he is the sort of person to pursue them. Hence the incontinent person is easily persuaded out of it while the intemperate person is not" (1999, 111; NE Book 7, Chapter 8, section 4).

> their frequent comorbidity. The evidence also suggests that apparently different types of repeated negative ideation, including worry in GAD [generalized anxiety disorder] and rumination in depression, have more in common and are more similar across disorders than is sometimes supposed. (Mathews and MacLeod 2005)

It is important to note that the cognitive biases specifically identified are focus biases[24] and recall biases, [25] which are both identified above as important causal factors in addictive behavior. Importantly, these biases disappear when emotional disorders are in remission (MacLeod and Mathews 1991). This data has given researchers reason to wonder whether attempts to address these cognitive biases would improve clinical outcomes.

In the case of *akrasia*, the kind of cognitive bias modification that should be effective given the cognitive account of *akrasia* that I have supplied, is as follows. The key to avoiding *akrasia* is attending to one's own better judgment. The *akrates* needs some form of cognitive bias modification that has the effect of combating the focus and recall biases that crowd attention to better judgment out of the global workspace. Such approaches have commonsense support. I am not the first to propose that such cognitive approaches are effective remedies for *akrasia*. Alfred Mele, in discussing what *enkrateia* (the opposite of *akrasia*) consists of, writes:

> An agent can, for example, keep clearly in mind, at the time of action, the reasons for doing the action which he judged best; he can refuse seriously to entertain "second thoughts" concerning matters about which he has just very carefully made up his mind; he can seek to add to his motivation for performing the action judged best by promising himself a reward (e.g., an expensive dinner) for successfully resisting temptation. (Mele, Self-Control, Action, and Belief 1985)[26]

---

24. For more evidence concerning the causal role of focus bias in emotional disorders like anxiety and depression, see (Mineka and Sutton 1992; Mathews and MacLeod 1985).

25. For more evidence concerning the causal role of recall bias in many emotional and other disorders, see (Blaney 1986).

26. Mele's view is not a fully worked out view of the motivational role of normative judgments, but his focus on specifically cognitive "therapies" is apropos. He cites Alston, "Self-Intervention and the Structure of Motivation" *The Self: Psychological and Philosophical Issues* ed. Mischel, Oxford: Blackwell, 1977, p.77 and Brandt, *A Theory of the Good and the Right* Oxford: Clarendon, 1979, pp. 111, 126-27, 333ff.

The first part of the sentence refers to keeping better judgment in the global workspace, or as in Ingrid's case, keeping unwanted stimuli out of it. The second part of the sentence refers to strategies like Alex's, though his reward is supporter approbation while failure carries a penalty. Commonsense approaches to behavior analysis and modification are not always accurate, but where careful study confirms them, we have that much more reason to rely on such approaches.

Earlier, I mentioned the relationship between decisional heuritics and cognitive bias. What is worth noting is that a heuristic is a passive thing, while metajudgment is active, and requires attentional resources (i.e., space in the global workspace of consciousness). It is also slower and more deliberate. Common remedies for attending to better judgment often feature a strategy of being more cognitively active than passive. Counting to ten before acting or speaking gives the agent opportunity to attend to metajudgment rather than acting out of anger or other impulse. Posting reminders to oneself where they will be seen during critical moments helps people to attend to factors that they at once consider most important and at the same time know they may neglect.

The success of some of these long-used attempts at cognitive bias modification is also observed in a more controlled setting. A recent study by Hoppitt, Matthews, Yiend, and Mackintosh (2010) examines the role of active training in cognitive bias modification. The study is designed to reveal the effect of active (as opposed to passive) training on modifying cognitive bias.

The study takes two groups of volunteers who are not disposed to anxiety, as measured by a standardized assessment. One group is given active cognitive bias modification, while the other group is given passive cognitive bias modification. In the active training, the subjects are given a scenario that is emotionally ambiguous until the last word of the scenario. For example:

> You have decided to go caving even though you feel nervous about being in such an enclosed space. You get to the caves before anyone else arrives. Going deep inside the first cave you realize you have completely lost your w—. (Hoppitt, et al. 2010, 75)

The framers of the study point out that such a scenario is emotionally ambiguous in the sense that the last word could sensibly be 'fear,' but supplying the first letter of the word 'way' resolves the ambiguity. The subject is then asked if they envision themselves feeling afraid in the cave.

The passive training group is supplied with the entire passage above, complete with the final word, and the sentence 'You are feeling afraid of being in the cave' appended to

the end of the original passage. Both groups are then given a filler task and then are both presented with an emotionally ambiguous passage such as:

> You are finding that your sight is worse than it was and despite the risks you decide to try an experimental laser surgery you've read about. Afterwards as the bandages are taken off your eyes, you realize that your life will be affected radically by the results (Hoppitt, et al. 2010, 75).

The point is to see if there is a difference between how the actively trained group and the passively trained group interprets the ambiguous passage. The study found a statistically significant difference in the tendency of the actively trained group versus the passively trained group to interpret the ambiguous passage negatively. Presumably if the active training were positively valenced instead of negatively as the study write-up indicates then the active training would have increased the tendency of the active training sample to interpret the ambiguous passage positively.

Interpreting ambiguous evidence as valenced in a particular way is evidence of cognitive bias. If there were no cognitive bias present, the subject would interpret the ambiguous evidence as ambiguous. What the results of this study seem to indicate is that actively engaging the cognitive faculties to interpret data and envision one's own emotional response has an observable causal effect on future responses. Active cognitive engagement is at the heart of cognitive bias modification.

The study is carefully crafted to isolate the effect of active cognitive training, but the study interestingly confirms a great many common platitudes about behavior modification. For example, some form of "visualizing success" is a staple in self-help guides and guides to personal and professional success. The idea is that when you actively visualize yourself acting, thinking, or deciding a certain way, you become more likely to act, think, and decide in that way.

The treatment of a focus bias, especially in cases of addiction, would then have a strong effect on determining whether the addict would refrain or relapse. Most of the work in modifying focus bias is in the context of treatments for anxiety disorders. Part and parcel of the anxiety disorder is focusing unduly on negative or threatening stimuli to the exclusion of positive or non-threatening stimuli. There are two ways of measuring anxiety: trait anxiety and state anxiety. Measures of state anxiety are measures of the degree to which a person is in an anxious state. Trait anxiety is a measure of the effect of anxiety-producing stimuli. A recent review of the literature concerning attentional bias modification indicates that "Attention Bias Modification Treatment produced a greater

effect on trait than state anxiety measures. This suggests that ABMT might target the more enduring aspects of anxiety" (Hakamata, et al. 2010).

The message is encouraging for the treatment of *akrasia* by means of treating the cognitive biases that are implicated in the *akrates*. If the anxiety sufferer can come to diminish attentional biases that select threatening stimuli to the exclusion of positive and neutral stimuli, then it stands to reason that the addicted *akrates* like Ingrid or Alex may train him or herself to focus on more stimuli in their environments other than alcohol-related or gaming-related stimuli.

A study by Lester and others, similar to the Hoppit et al. study described above, but with a broader scope, details some strategies for cognitive bias modification designed to broadly treat anxiety and depression. What should strike the reader about their descriptions is that they are much more pedestrian in nature than the clinically impressive sounding phrase 'cognitive bias modification therapy' would suggest. It is a case in which at least some aspects of our common folk psychology have some empirical verification in a carefully controlled setting.

A sampling of the cognitive biases and their modification strategies are as follows (Lester, et al. 2011, 300):

Cognitive Error	Definition	Clinical Example	Example Modification Item
Selective Abstraction	Focusing on a detail taken out of context, while ignoring other more salient features of the situation and conceptualizing the whole experience on the basis of this fragment	A recent graduate begins a new position and is eager to make friends with their colleagues. They ask their new colleagues whether they would like to join them for a drink after work and 2 people accept their offer. They focus on the fact that some people declined and think this means they aren't liked rather than being pleased that some of their colleagues are keen to socialize.	You have started a new job and hope to be friends with your colleagues. At the end of your first day you ask whether people would like to go for a drink and 2 people offer to come out with you. You think this means you have probably been rejected/ accepted Have you failed to make friends?
Dichotomous Thinking	Tendency to place all experiences in one of two opposite categories, e.g. flawless or defective rather than viewing them as existing on a continuum. In describing oneself, the extreme negative categorization is selected	You've been trying to diet but you've eaten a few sweets over the weekend. You tell yourself that you can never control yourself and that all your dieting and jogging over the whole week have gone down the drain.	You have been on a really strict diet for a few weeks and have totally cut out sweet things. However you couldn't resist a piece of cake on your friend's birthday. You think your attempts at dieting have been… futile/disciplined Have you completely failed in your attempts to diet?

Notice the overlap between the cognitive errors described in this table and cognitive errors involved in classic examples of *akrasia* discussed throughout this work and other philosophical discourse on *akrasia*. The examples in the Lester et al. study are tailored to anxiety and depression, but consider different ways of fitting the definitions supplied.

Ingrid is at a party, and there is alcohol present, and several people near her are having an alcoholic drink. Ingrid focuses unduly on these examples and becomes anxious

that everybody else is drinking, and she feels a great deal of social pressure that crowds out her resolve to stay on the wagon. Now imagine that a close friend is next to her to apply cognitive bias modification treatment. This interlocutor points out all of the people who are not drinking alcohol, and asks probing questions of Ingrid, asking whether she really believes that anyone notices or cares whether or not she has a drink. This line of questioning and pointing out of external stimuli actively engage Ingrid's cognitive faculties and gives her a greater chance to attend to her better judgment of abstinence.[27]

Consider now any dieter at some stage of the process, who can be accused of dichotomous thinking with rather little modification of the above example. The dieter, though judging that it would be better to avoid the sweets than to indulge in them, recalls *akrasia* in his recent past, and considers his diet irrevocably lost. He indulges in the sweets, contravening his better judgment while making it even easier to continue indulging in the sweets. Again, an interlocutor could actively engage his cognitive faculties with probing questions about the real effectiveness of dieting and the comparative effectiveness of indulging less as opposed to more. Again, this would have the effect of not only allowing better judgment to prevail in this case, but (in accord with the evidence from the Hoppitt study) makes it more likely to prevail in similar circumstances in the near future.

The success of these strategies for cognitive bias (and therefore behavior) modification is also confirmed by Lester et al. In their words, "Cognitive Error Modification was capable of inducing systematic group differences in how hypothetical events were perceived in both a healthy and vulnerable sample" (305).

Of course, strategies for anti-*akratic* cognitive bias modification need not necessarily involve an interlocutor. Controlling one's environment (as in Odyssean self-control), setting reminders for oneself in places that they will likely be seen (being one's own interlocutor), habituating active engagement of cognition and metacognition (repetition of slogans, mottos, or using the 'count to ten' strategy) are all examples of cognitive bias modification therapy that do not require a therapist.

I hope I have not belabored the point, but what I have been arguing is that the right way to reform the *akrates* is to focus on the cognitive aspects of the *akrates* rather than on their desires. If *akrasia* involves cognitive bias, and if the difference between being *akratic* and not being *akratic* is actuated on the modification of cognitive states, then

27. Consider this from Aristotle: "For some people are like those who do not get tickled themselves if they tickle someone else first; if they see and notice something in advance, and rouse themselves and their rational calculation, they are not overcome by feelings, no matter whether something is pleasant or painful" (Nicomachean Ethics 1999, 110; Book 7, Chapter 7, Section 8).

this is good reason to believe that the cognitive account of *akrasia* is the right account. If the cognitive account is the right account, that indicates that normative judgments, understood as cognitive states, play a significant role in motivation and action. The evidence I have gone to such lengths describing is at odds with the picture of *akrasia* painted by the Humean. For the Humean, you can judge and cogitate all you like, but unless you have the appropriate desires, your behavior doesn't change. The evidence indicates that cognitive states (which include normative judgments) have a much more significant role than the Humean perspective allows in motivation and action.

Bibliography

Alcoholics Anonymous. 2001. *The story of how many thousands of men and women have recovered from alcoholism* (4th ed.). Alcoholics Anonymous World Services, Inc.

Aristotle. 1999. *Nicomachean Ethics*. Translated by Terence Irwin. Indianapolis: Hackett.

Baars, Bernard. 2003. "The Global Brainweb: An Update on Global Workspace Theory." *Science and Consciousness Review* (October).

Blaney, Paul. 1986. "Affect and Memory: A Review." *Psychological Bulletin* 99 (2): 229–246.

Boyer, Morten, and Mark Dickerson. 2003. "Attentional bias and Addictive behavior: Automaticity in a gambling-specific modified Stroop task." *Addiction* 98 (1): 61–70.

Campbell, William. 2003. "Addiction: A Disease of Volition Caused by a Cognitive Impairment." *The Canadian Journal of Psychiatry* 48 (10), 669–674.

Cosmides, Leda, and John Tooby. 1994. "Better than Rational: Evolutionary Psychology and the Invisible Hand." *American Economic Review* 84 (2): 327–332.

Cox, W. Miles, Lee Hogan, Marc Kristian, and Julian Race. 2002. "Alcohol attentional bias as a predictor of alcohol abusers' treatment outcomes." *Drug and Alcohol Dependence* 68 (3): 237–243.

Davidson, Donald. 2001. "How is Weakness of the Will Possible?" In *Essays on Actions and Events*, 21–42. Oxford: Clarendon Press.

Davidson, Donald. 2004. "Paradoxes of Irrationality." In *Problems of Rationality*, 169–187. Oxford: Oxford University Press.

Decker, Seamus, and Jessica Gay. 2011. "Cognitive-bias toward gaming-related words and disinhibition in World of Warcraft gamers." *Computers in Human Behavior* 27 (2): 798–810.

Gigerenzer, Gerd, and Daniel Goldstein. 1996. "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review* 103 (4): 650–669.

Gigerenzer, Gerd, and Peter Todd. 1999. *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Hakamata, Yuko, and Shmuel Lissek, Yair Bar-Haim, Jennifer C. Britton, Nathan A. Fox, Ellen Leibenluft, Monique Ernst, Daniel S. Pine. 2010. "Attention Bias Modification Treatment: A Meta-Analysis Toward the Establishment of Novel Treatment for Anxiety." *Biological Psychiatry* 68 (11): 982–990.

Hardcastle, Valerie. 2003. "Life at the borders: habits, addictions, and self-control." *Journal of Experimental & Theoretical Artificial Intelligence* 15 (2): 243–253.

Hare, Richard. 1963. *Freedom and Reason*. New York: Oxford University Press.

Hester, Robert, Veronica Dixon, and Hugh Garavan. 2006. "A Consistent Attentional Bias for Drug-Related Material in Active Cocaine Users Across Word and Picture Versions of the Emotional Stroop Task." *Drug and Aclohol Dependence*, 81 (3): 251–257.

Hoppitt, Laura, Andrew Mathews, Jenny Yiend, and Bundy Mackintosh. 2010. "Cognitive Bias Modification: The Critical Role of Active Training in Modifying Emotional Responses." *Behavior Therapy* 41 (1): 73–81.

Janes, Amy, Diego Pizzagalli, Sarah Richardt, Blaise deB. Frederick, Sarah Chuzi, Gladys Pachas, Melissa A. Culhane, Avram Holmes, Maurizio Fava, A. Eden Evins, and Marc Kaufman. 2010. "Brain reactivity to smoking cues predicts ability to maintain tobacco abstinence." *Biological Psychiatry* 67 (8): 722–729.

Kalis, Annemarie, Andreas Mojzisch, T. Schweizer, and Stefan Kaiser. 2008. "Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective." *Cognitive, Affective, & Behavioral Neuroscience* 8 (4): 402–417.

Lester, Kathryn, Andrew Mathews, Phil Davison, Jennifer Burgess, and Jenny Yiend. 2011. "Modifying cognitive errors promotes cognitive well being: A new approach to bias modification." *Journal of Behavior Therapy and Experimental Psychiatry* 42 (3): 298–308.

Liu, Shijing, Scott Lane, Joy Schmitz, Andrew Waters, Kathryn Cunningham, and F. Gerard Moeller. 2011. "Relationship between attentional bias to cocaine-related stimuli and impulsivity in cocaine-dependent subjects." *The American Journal of Drug and Alcohol Abuse* 37 (2): 117–122.

MacLeod, Colin, and Andrew Mathews. 1991. "Cognitive Experimental Approaches to the Emotional Disorders." In *Handbook of Behavior Therapy and Psychological Science,* edited by Paul Martin, 116–150. New York: Pergamon Press.

Mathews, Andrew, and Colin MacLeod. 1985. "Selective processing of threat cues in anxiety states." *Behavior Research and Therapy* 23 (5): 563–569.

Mathews, Andrew, and Colin MacLeod. 2005. "Cognitive Vulnerability to Emotional Disorders." *Annual Review of Clinical Psychology* 1: 167–195.

Mazas, Carlos, Peter Finn and Joseph Steinmetz. 2000. "Decision-Making Biases, Antisocial Personality, and Early-Onset Alcoholism." *Alcoholism: Clinical and Experimental Research* 24 (7): 1036–1040.

Mele, Alfred. 1985. "Self-Control, Action, and Belief." *American Philosophical Quarterly* 22 (2): 169–176.

Mele, Alfred. 1987. *Irrationality.* New York: Oxford University Press.

Mineka, S., & Sutton, S. K. 1992, January. Cognitive Biases and the Emotional Disorders. *Psychological Science* 3 (1): 65–69.

Nijs, Ilse, Peter Muris, Anja Euser, and Ingmar Franken. 2010. "Differences in attention to food and food intake between overweight/obese females and normal-weight females under conditions of hunger and satiety." *Appetite* 54 (2): 243–254.

Pinker, Steven. 2011. *The Better Angels of Our Nature.* New York: Penguin.

Smith, Michael. 2003. "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion." In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 17–38. Oxford: Clarendon Press.

Stocker, Michael. 1979. "Desiring the Bad: An Essay in Moral Psychology." *Journal of Philosophy* 76 (12): 738–753.

Vadhan, Nehal, Kenneth Carpenter, Marc Copersino, Carl Hart, Richard Foltin, and Edward Nunes. 2007. "Attentional Bias towards Cocaine Related Stimuli: Relationship to Treatment-Seeking for Cocaine Dependence." *American Journal of Drug and Alcohol Abuse* 33 (5): 727–736.

Watson, Gary. 1977. "Skepticism About Weakness of Will." *The Philosophical Review* 86 (3): 316–339.

Wedgwood, Ralph. 2003. "Choosing Rationally and Choosing Correctly." In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 201–229. Oxford: Clarendon Press.

Wedgwood, Ralph. 2007. *The Nature of Normativity.* Oxford: Clarendon Press.

Journal of Cognition and Neuroethics

Strong Emergence and Mental Causation in Gadamer's *Truth and Method*

Matthew E. Johnson
Institute for Christian Studies

Biography

Matthew E. Johnson is completing a Master of Arts in Philosophy at the Institute for Christian Studies in Toronto, Canada, specializing in Aesthetics, Hermeneutics, and Discourse. His research interests include the intersection of Continental philosophy, philosophy of mind, and American pragmatism, and his thesis is titled *Liberating Emergence: Human Dependence and Autonomy in Emergentism, Hermeneutics, and Pragmatism*.

Publication Details

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

Citation

Johnson, Matthew E. 2014. "Strong Emergence and Mental Causation in Gadamer's *Truth and Method*" *Journal of Cognition and Neuroethics* 2 (2): 31–50.

Strong Emergence and Mental Causation in Gadamer's *Truth and Method*

Matthew E. Johnson

Abstract

The arc of Hans-Georg Gadamer's *Truth and Method* introduces a difficult tension between (1) the way in which the human person is fundamentally dependent upon convention and, more broadly, historical and environmental situatedness, for existence and (2) the way in which human action is not fully pre-determined by this dependence. Throughout his hermeneutic ontology, Gadamer maintains that the human being is free in a legitimate sense. The boundaries of historicity and language simultaneously limit and enable the ability to come to a self-understanding that leads to novel interaction with the world. Because Gadamer strongly resists Cartesian dualism, he describes the human person's ability to resist the causal pressures of the environment in a way that maintains both the situatedness and the freedom of the human person. As a result, his hermeneutic ontology, with its development of the concept of play, the hermeneutic circle, and the linguistic structure of hermeneutic experience, bears a certain resemblance to concepts central to strong emergentism. As a means by which it is possible to account for both the full embeddedness of an emergent while maintaining its novelty and causal efficacy with respect to its originary system, strong emergentism provides tools with which to analyze and clarify how *Truth and Method*'s post-Kantian and post-Cartesian position retains and develops a sense of legitimate free will for Dasein within the boundaries of historical and environmental situatedness.

Keywords

Hermeneutics, Gadamer, emergence, free will, personhood, mind, dependence, autonomy

Introduction: The Hermeneutics of Brick and Blanket

For geniuses with IQs above a certain threshold (somewhere around 130), a higher IQ is essentially less useful than a brick when it comes to predicting the person's capacity to succeed in the real world. Canadian journalist Malcolm Gladwell's description of the limits of the typical IQ test's ability to predict success in his book *Outliers* illustrates an important feature of the relevance of a strongly emergent conception of human cognition to the hermeneutic philosophy of Hans-Georg Gadamer in his landmark work *Truth and Method*. Gladwell describes an alternative kind of test called a "divergence test," which he claims is a much more accurate predictor (2008, 90). The test involves the creative interpretation of (1) a brick and (2) a blanket; that is, the test-taker is given a limited amount of time to write down as many uses as possible for each. This measures the test-taker's ability to think *creatively*, as opposed to an IQ test's measure of only analytical

intelligence. For example, one particularly creative and clever test-taker interpreted the brick in this way:

> (Brick). To break windows for robbery, to determine depth of wells, to use as ammunition, as pendulum, to practice carving, wall building, to demonstrate Archimedes' Principle, as part of abstract sculpture, costh, ballast, weight for dropping things in river, etc., as a hammer, keep door open, footwiper, use as rubble for path filling, chock, weight on scale, to prop up wobbly table, paperweight, as fire- hearth, to block up rabbit hole. (Gladwell 2008, 88)

This test-taker likely had a similar range of responses for the uses of a blanket and, because of his creativity, probably scored quite high on the divergence test. Though most of us lie well below the genius IQ threshold, this divergence test is an excellent example of an idea that Hans-Georg Gadamer likely had in mind when he laid out his hermeneutic ontology in *Truth and Method*. The divergence test effectively measures the test-taker's ability to *interpret* and *understand* the brick and the blanket and to take a critical stance that does not simply conform to the conventional use of these objects, an ability that involves creativity and free thinking.

The arc of *Truth and Method* introduces a difficult tension between the way in which the human person is fundamentally dependent upon convention and, more broadly, historical and environmental situatedness, for existence on the one side and the way in which human action is not fully pre-determined by this dependence on the other. Throughout his hermeneutic ontology, Gadamer maintains that the human being is free in a legitimate sense. The boundaries of historicity and language simultaneously limit and enable the ability to come to a self-understanding that leads to novel and innovative interaction with the world. Gadamer designates this ability as "*freedom from environment*" (2004 [1989], 441). Because Gadamer strongly resists Cartesian dualism, he must explain the origin of this ability to resist the causal pressures of the environment in a way that maintains both the situated dependence and the autonomous freedom of the human person. As a result, his hermeneutic ontology, with its development of the concept of play, the hermeneutic circle, and the linguistic structure of hermeneutic experience, parallels insights drawn from strong emergentism.

In order to clarify the way in which Gadamer negotiates this complex course through an ontology of human dependence to the freedom of the human being, I suggest that his line of reasoning can be helpfully illuminated in terms of strong emergentism, which seeks to answer similar concerns. As a means by which it is possible to account for both

the full embeddedness of an emergent while maintaining its novelty and causal efficacy with respect to its originary system, emergentism provides us with tools to analyze and clarify how *Truth and Method*'s post-Kantian and post-Cartesian position retains and develops a sense of legitimate free will for Dasein within the boundaries of historical and environmental situatedness.

The method of analysis that follows will involve the exposition of Gadamer's hermeneutic ontology in *Truth and Method* translated into emergentist terms, showing how the two frameworks of thought naturally converge on similar conclusions. This compelling convergence will both clarify Gadamer's ontology and pave the way for a compelling case for strong emergence reinforced by Gadamer's hermeneutic ontology in *Truth and Method*. The paper will be divided into three parts. The first will briefly outline the relevant concerns in the emergentism literature, and the second will develop Gadamer's concepts of play and the hermeneutic circle in terms of emergentism. The final part will draw out the implications for Gadamer's strong emergentism for mental causation and, ultimately, a case for the possibility of free will based on emergence, which will be developed as a breed of downward causation.

Setting the Stage

To begin, because Gadamer stands on the shoulders of Martin Heidegger, it is essential to explicate the Heideggerian basis of certain relevant aspects of Gadamer's project. In *Being and Time*, Heidegger defines "understanding" as "*the existential being [Sein] of the ownmost potentiality of being of Dasein itself in such a way that this being [Sein] discloses in itself what its very being is about*" (1996, 144). For Heidegger, understanding always involves sifting through the various possibilities for one's future activity from moment to moment. These possibilities present themselves always and only in terms of a world. Sifting in such a way allows the various possibilities to be interpreted in terms of "service*ability*, us*ability*, detriment*ality*" (Heidegger 1996, 144). The process of developing and arriving at an understanding of the availability and quality of possibilities, then, is what Heidegger calls "interpretation." So when Dasein interprets, it comes to understand *itself* in terms of the possibilities available for its activity, always and only in relation to the object(s) of interpretation. Understanding and interpretation, from a Heideggerian point of view, move beyond Cartesian dualism and the Kantian subject-object schema, placing the being of Dasein *in* the act of interpretation itself in such a way that Dasein exists as external to itself (so to speak). Understanding, then, is not a matter

of collecting knowledge; it is a fundamental mode of Dasein's being, which Donatella Di Cesare describes as being as close to us and as inescapable as breathing (2013, 38).

Showing his Heideggerian hand in the foreword to the second edition of *Truth and Method*, Gadamer notes that his use of the term *hermeneutics* "denotes the basic being-in-motion of Dasein that constitutes its finitude and historicity, and hence embraces the whole of its experience of the world" (2004 [1989], xxvii). For Gadamer, hermeneutics is not simply the interpretation of texts; it is a fundamental mode of being in which a person exists in the world. The path Gadamer takes to develop this claim in *Truth and Method* is through an ontology of human dependence upon the world in which both "subject" and "object" have their being only in a relationship to each other (in *presentation* and *interpretation*), a relationship that always runs both ways between them. Gadamer identifies this relationship as "play" and establishes it as the dynamic that enables interpretation in general.[1] For Gadamer, interpretation is *the* fundamental mode of being of the human being. So it is not that we, as transcendental subjects, deign to enter into a relationship of interpretation, but rather that we are always already relationally involved in a historical world and express our being through the dynamics of interpretation. *Tradition* and *prejudice*, in Gadamerian terms, fill in the content of the human person's historical situatedness and constitute the way in which we inextricably belong to history (2004 [1989], 278). In this way, as Stefano Marino explains, Gadamer's hermeneutic ontology is a re-conception of what it means to be human that converts philosophical hermeneutics into a practical philosophy, taking into account the way in which humans are always already situated and involved in a complex natural, social, and historical environment (2011, 217). Once again aligning with Heidegger in an attempt to make clear the structure of interpretation, Gadamer says: "Everything that makes possible and limits Dasein's projection ineluctably precedes it" (2004 [1989], 254). Dasein has no being, let alone self-understanding, apart from its being already embedded in and relationally connected to a world. This embeddedness limits what Dasein can do and consider doing but at the same time constitutes the array of possibilities available to interpreting Dasein.

At this point, a problem is introduced that requires resolution. In Gadamer's hermeneutic ontology, there is heavy emphasis on the historical constitution of Dasein

1. Though in *Truth and Method*, Gadamer's explicit use of the concept of play is used mostly in reference to his ontology of the work of art, along with Monica Vilhauer (2010, xiii-xiv), I suggest that the concept of play is a foundational concept for his hermeneutic ontology in general. As such, it establishes a foundation upon which to build his more broadly scoped ideas.

and the way in which Dasein gets caught up in interpreting and interacting with the world. This poses a challenge to maintaining a robust sense of Dasein's free will: if Dasein is fully pre-constituted by its historicity and embeddedness in a world, is it possible for Dasein to have the ability to choose between possibilities in a way that is not already predetermined by its situatedness? Are legitimate possibilities ever presented to Dasein as having equipotentiality, or is every projection fully determined by Dasein's historical constitution? Can Gadamer maintain that Dasein is profoundly and ontologically dependent on historical situatedness and still account for any sense of legitimate freedom of the human being without reverting back to Cartesian dualism? I suggest that Gadamer's account of freedom in the hermeneutic ontology of *Truth and Method* successfully deals with these issues, and, in doing so, crosses paths with strong emergentism.

I. Emergence: A Third Way

Much like Gadamer's hermeneutic ontology, the renewed interest in emergentism in philosophy of mind, philosophy of science, and philosophy of religion has been motivated in large part by the ongoing failure of the scientific pursuit of a complete explanatory reduction of the universe to a single set of laws (i.e., physics) (Clayton 2006, 1). With the apparent futility of this project on one side and the well-established uneasiness with Cartesian dualism on the other,[2] philosopher of religion James W. Haag endorses emergentism thus: "Emergentism, by occupying the gap between reductive Physicalism and substance Dualism, provides a viable worldview" (Haag 2008, 12). As such, emergentism allows for the possibility that a phenomenon or entity can be "at once *grounded in* and yet *emergent from* the underlying material structure with which it is associated" (O'Connor 1994, 91), thereby making way for an antireductionistic means by which to describe the universe without either falling into substance dualism or writing off the legitimacy of the physical sciences. Walking this tightrope, Michael Silberstein

2. One of René Descartes' most significant contributions to Western philosophy consisted of a distinction between the human body and the mind, between *res extensa* ("thing that is extended physically") and *res cogitans* ("thing that thinks"). This dualism led to what has come to be known as the mind-body problem, a problem which many Western philosophers aim either to solve, resolve, or dissolve. However, before blaming Descartes for all the problems in Western philosophy, it should be noted that Descartes himself offered what he considered to be a solution to the mind-body problem. Mark A. Bedau notes that Descartes developed an account of how *res cogitans* and *res extensa* interact, even though they are two types of substances, based on the idea that *res cogitans* was emergent from a bodily organ (Bedau 1986). The mind-body problem was extended and complicated by post-Cartesians who took up Descartes' problem without accepting his proposed solution.

argues for a version of strong emergence (which he refers to as "ontological emergence"), describing the position as a way to maintain compatibility between metaphysical monism and ontological pluralism (2006, 206). The appeal of emergence theory lies in its explanatory power as a legitimate third way that acknowledges the complex relations in the world and the irreducibility of these relations to a single vocabulary or substance (e.g., elementary particles).

Broadly, an emergent property may be defined as follows: "a property P is *novel* in x if x has P, and there are no determinates P' of the same determinable as P, such that any constituents of x have P'" (Spencer-Smith 1995, 117). That is, an emergent (P) shows up in, or as a result of, a system (x) and cannot be reduced to the components of x; the origin of P requires the entire system and cannot be traced back to individual components. However novel an emergent might be, it is always a property of the system as a whole, never of simple component parts (Georgiou 2003, 240). Stuart Kauffman gives this principle a temporal spin, offering emergence as an alternative to simplistic Newtonian physics. Emergence, for Kauffman, is marked by a novelty that is not time reversible. Whereas according to Newton's laws, an object traveling in one direction can retrace its steps and remain the same object, a human being's experience of being-in-the-world (for example) creates a state of constant flux from one moment to the next in which a human consciousness, as emergent from its being-in-the-world, cannot remain precisely the same through time. "[A]s Humpty Dumpty famously discovered," writes Kauffman, "we are not time reversible. Neither is the world around us" (Kauffman 2008, 13).

So a theory of emergence must accept some variation of the basic thesis on the origin of novel emergents and will typically grapple with at least four additional criteria, such as those identified by Philip Clayton: (1) ontological monism; (2) property emergence; (3) the irreducibility of the emergence; and (4) downward causation (2006, 2). These four criteria, however, are far from representing a consensus in the literature. Rather, they are four of the primary points of contention among emergentists. Of the four, however, downward causation is perhaps the most polarizing, leading to a stark bifurcation of the field. Disagreement on downward causation (that is, the idea that an emergent "exert causal influence 'downward' to affect the processes at a lower basal level" [Kim 2006, 198]) birthed a difference between "strong" and "weak" emergence. A "strong" position will claim the legitimacy of downward causation (often simultaneously challenging ontological monism). More formally, strong emergence consists in the following:

Property P is an emergent property of a (micrologically-complex) object O iff:

1. P supervenes on properties of the parts of O;

2. P is not had by any of the object's parts;

3. P is distinct from any structural property of O; and

4. P has direct ('downward') determinative influence on the pattern of behavior involving O's parts (O'Connor 1994, 98).

On the other hand, a "weak" position will deny the actual existence of an emergent as an entity or property capable of downward causation, limiting emergence to an explanatory shortcut, useful for describing the behaviour complex wholes.

Strong emergence, with its inclusion of downward causation, was central to early evolutionary theory, particularly since the work of Conway Lloyd Morgan, who observed that evolution consists of a series of emergent steps, each of which introduce something new to the evolutionary progression that changes its course (Morgan 1927, 1). Explaining the emergence of life from this perspective, Kim Sterelny calls one point at which downward causation begins to occur the "organism threshold." Above this threshold, "natural selection typically acts directly on organisms and indirectly on [self-replicating proteins]" (Sterelny 2001, 23). The organism threshold marks the point at which the behaviour of the whole organism directly affects how the genetic material is selected. In this view, organisms are characterized by "emergent properties not found at the level of their molecular components" (Baetu 2012, 434). Even among strong emergentists, however, there is considerable disagreement about how exactly downward causation works. Debates about downward causation are inextricably linked to concerns central to philosophy of mind such as intentionality and mental causation, as well as to the nature of the mind in general. Many strong emergentists maintain that mental causation is a clear example of downward causation (Kim 2006, 198), giving these theorists a fresh framework within which to analyze human behaviour and individuality. Thus, the richness of emergentism's contributions to philosophy of mind creates fertile ground for new positions on the question of determinism versus the free will of emergents (in particular, human persons).

In recent years, many emergentists have begun to lose faith in the weak emergentist's loyalty to the project of scientific reductionism. Whereas the British Emergentists of the late 19th and early 20th centuries, heavily influenced the work of J. S. Mill, G. H. Lewes and others, welcomed a strong conception of emergence, the later century's intense optimism in the project of the scientific reductionism made way for weak emergence to become something of an orthodoxy in the field by the late 20th century (Haag 2008, 43–44). Recent years, however, have seen a reawakening of openness to strong emergence in the literature

(Clayton 2006, 27), and with it, a new philosophical interest in an emergentist take on what may be considered irreducible aspects of the world, including human historicity, life as such, and human social reality (Kauffman 2008, x). While this does not necessitate the abandonment of faith in the pursuits of the physical sciences, it does signify movement toward a wider and more nuanced understanding of the universe and of human existence. Not only this, but weak emergentism's insistence on ontological reductionism results in a truncated view of human freedom that does not adequately account for the perception of our own agency that we experience every day. Whereas weak emergence tends to go hand in hand with the determinism of a reductionist ontology, strong emergence remains open to the possibility of human free will. I suggest that this increasing openness to strong emergence is movement closer to Gadamer's hermeneutic ontology. Thus, having traced some of the core concerns of emergence theory, a foray into Gadamer's *Truth and Method* will prove fruitful in the defense of strong emergence. As I will show, Gadamer converges on a theory of strong emergence with regard to human consciousness that maintains a non-reductive view of human free will and downward (mental) causation.

II. Gadamer's Strong Emergentism

There is a recent precedent for connecting Gadamer's hermeneutic ontology to positions related to emergentism in philosophy of mind. In his article "The Source of the Subjective," Bjørn Ramberg argues for the juxtaposition of Gadamer's hermeneutic ontology against an analysis of intentionality that is based on the understanding that the mind "exists *as* a system of relations" between the human person and environment (1997, 467). Although Ramberg does not explicitly delve into emergentism for support of this thesis, his suggestions are deeply compatible with the core intuitions of strong emergentism, and the connections he makes in this article neatly pave the way for this juxtaposition to be developed.

Play and the Conditions for Emergence

To unpack Gadamer's compatibility with strong emergentism, I will begin with an exposition of his concept of *play*, juxtaposed against an emergentist ontology. Whereas much of the secondary literature on Gadamer deals with the concept of play primarily in relation to his ontology of the work of art, Monica Vilhauer suggests that therein lies the key to understanding Gadamer's philosophical hermeneutics. She suggests that this key concept

> elucidates the very process of understanding *in general*—the understanding which stretches through all our hermeneutic experience, including encounters with art, with text, with tradition in all its forms, with others in dialogue, and which even constitutes our very mode of being-in-the-world. (Vilhauer 2010, xiii-xiv)

Considering that the Heideggerian sense of *understanding* and *interpretation* describes a fundamental way in which Dasein is oriented toward the world, the concept of play, as the dynamic that animates both understanding and interpretation, brings Gadamer's hermeneutic ontology into focus.

Gadamer considers play to be the actual mode of being of a work of art (2004 [1989], 102). That is, a work of art only has its being in *being played* by, and thereby presenting itself to, an observer or participant. However, this self-presentation goes both ways. Thus, play is not only the mode of being of the art object but is also an occasion in which the human person engages in interpretation, which, understood in a Heideggerian sense, also constitutes the being of the person: "in spending oneself on the task of the game, one is in fact playing oneself out. The self-presentation of the game involves the player's achieving, as it were, his own self-presentation by playing—i.e. presenting—something" (Ibid., 108). The player engages with the game or the work of art as the "space" in which to project and imagine possibilities (Ibid., 250). This projection is what Gadamer identifies as one's "ecstatic self-forgetfulness," which he suggests paradoxically "corresponds to [one's] continuity with [oneself]" (Ibid., 124). Therefore, the human person, according to Gadamer, only exists as always already engaged in play, that is, always extended into (and, I suggest, emergent from) relationships with the world. There is an emergence that takes place here through the mode of self-understanding by which one who understands something in the world (i.e., "projects oneself upon his possibilities" [Ibid., 251]) understands oneself. This interpretation of Gadamer's use of the Heideggerian concept of understanding allows Dasein to be constituted by components that do not exhaust its being; Dasein emerges out of a system of interrelationships in the world within which it is caught up in play, and Dasein is only intelligible to itself in terms of these pre-established interrelationships.

The concept of play is a key component to illuminating an interpretation of Gadamer's hermeneutic ontology that is amenable to a strong emergentist account of consciousness. Vilhauer suggests that Gadamer, through the concept of play, offers a solution to the mind-body problem, simultaneously challenging the Kantian view that a person is a subject as opposed to objects and the Cartesian view that the mind is a distinct kind

of entity from the body. Rather, as a result of the relational and ontological significance of play, Vilhauer's Gadamer offers a view in which the human person "is a being that is primordially in contact with the world of meaningful things and people, apart from which this thing cannot exist" (Vilhaeur 2010, 112). Similarly, developing a juxtaposition of hermeneutics and intentionality, Bjørn Ramberg maintains that mental properties, in an externalist (and, I might add, emergentist) view of mentality and intentionality,

> are not autonomous, intrinsic features of some entity; they are essentially relational. They are individuated, and constituted (in part) by objects beyond the subject or person. A person's mental properties are a system of relationships between the person and her environment. (Ramberg 1997, 467)

In the same way that a strong emergentist is able to consider mental activity as a process of interaction between "mutually embedding and embedded systems, tightly interconnected on multiple levels" (Silberstein 2006, 208), rather than an inner quality of an individual, Gadamer also views the human person as constituted by this very dynamic, which he calls hermeneutics (i.e., "the basic being-in-motion of Dasein" [Gadamer 2004 [1989], xxvii]). Such a view of the human person is summed up in the words of Warren Brown's (and John Dewey's) insistence that "mind" should be understood as a verb and not a noun (Brown 2007, 200). In the same way, for Gadamer, the concept of play illustrates a view of consciousness as always in motion and always caught up in the world.

The Hermeneutic Circle and the Dynamics of Emergence

Using the image of the hermeneutic circle, fortified by a nuanced understanding of *tradition* and *prejudice* as essential components of the human person's fundamental constitution, Gadamer moves from the concept of play to conceiving of the human person as historically embedded. If self-presentation in play corresponds roughly to Heidegger's notion of projection, Gadamer's conceptions of tradition and prejudice correspond to Heidegger's notion of heritage. Gadamer affirms and alludes to something similar to Heideggerian heritage as a way to describe the human person as belonging to history (2004 [1989], 278): "Everything that makes possible and limits Dasein's projection ineluctably precedes it" (Ibid., 254). For Gadamer, as for Heidegger, the emergence of a self that is capable of authentic projection is not possible apart from being already conditioned by a historical situation. One's historical constitution consists in a set of prejudices that makes possible all understanding and serves to direct and orient inquiry (Ramberg 1997, 460-461). A human person always begins with a set of prejudices, or in more Heideggerian

terms "fore-conceptions" (Gadamer 2004 [1989], 269), that are primordial as a result of being always and already immersed in a tradition (i.e., a heritage). However, for Gadamer, these prejudices initiate a process of ongoing interpretation. When these prejudices are challenged by the "text" (i.e., the object of interpretation), they cause the interpreter to be "pulled up short" (Ibid., 270) by it, and the interpreter is able to replace previous prejudices with new, more appropriate interpretations (Ibid., 269). So prejudices, for Gadamer, are more than just biases; they "*constitute the historical reality of [a human person's] being*" (Ibid., 278).

The process of testing prejudices against objects of interpretation is suitably deemed the "hermeneutic circle," which I suggest illustrates the process by which an individual *emerges* as a free individual out of thrownness: "The circle…is neither subjective nor objective, but describes understanding as the interplay of the movement of tradition and the movement of the interpreter" (Gadamer 2004 [1989], 293). To recall Stuart Kauffman's idea that an emergent emerges in a way that is not time-reversible, Gadamer's hermeneutic circle suggests the same. The movement of interpretation from prejudice to reformulation necessitates that understanding, which constitutes the very being of Dasein, is in a constant state of flux. A new understanding of the world cannot be erased without tampering with Dasein's primordial being-in-the-world. In this way, the emergence of Dasein from its heritage is not time-reversible but is diachronic and profoundly historically contingent.

From an emergentist's perspective, this means that there is no mind or consciousness at all apart from historical embeddedness. To suggest otherwise falls into a dualistic idea that the mind leads a separate existence from historical and physical embeddedness, a position that strong emergentists (and Gadamer) reject. This amounts to the idea that a brain in a vat can never have the same experiences as an identical brain in a body embedded in a historical situation (Silberstein 2006, 211). In fact, apart from this embeddedness, there is no possibility of a hermeneutic circle, and therefore, no possibility of the understanding that is constitutive of consciousness itself.

Where the hermeneutic circle does occur, however, an emergence takes place. This emergence is marked by the ability of the emergent (Dasein) to "foreground" a prejudice. In Gadamer's terminology, "Foregrounding (abheben) a prejudice clearly requires suspending its validity for us" (2004 [1989], 298). The ability to take such a critical stance on a prejudice requires the interpreter to resist the pressure it exerts. The human person, when actively involved in the world as an interpreter, emerges through the hermeneutic circle as something *more* than simply a bundle of prejudices, pre-determined by historical and social situatedness, even though these prejudices ground its being. So while a

human person only comes into being, so to speak, when engaged with the world in a play relationship marked by the hermeneutic circle, the being that has come into being emerges as more than the sum of the component parts that conditioned the possibility for its existence.

For Gadamer, the "more" that emerges out of the movement of the hermeneutic circle is marked by a "state of new intellectual freedom" (2004 [1989], 251). He goes on to explain that when a person comes to an understanding, this "implies the general possibility of interpreting, of seeing connections, of drawing conclusions" (Ibid., 251). I suggest that the freedom Gadamer attributes to Dasein (as an emergent) thus takes the form of *critical creativity* with respect to its thrownness, which means that such a person is able to prevent novel questions and lines of inquiry from being covered over by inherited prejudices (Ibid., 361). Such an ability is marked by the possibility of "taking a critical stance with regard to every convention" (Ibid., 551), which opens up the possibility of freedom from the pressures of these conventions. Persons who *understand* are thus not completely pre-determined by their inescapable historical constitution and their belonging to a tradition. The ability to engage critically with convention (i.e., one's thrownness into a tradition) results from tracing the path of the hermeneutic circle to arrive at a new self-understanding: "This cultivated understanding and self-understanding constitutes for us a newfound *freedom* in which we feel at home in what may have previously been strange and posed a limitation for us" (Vilhauer 2010, 65). In this way, the hermeneutic circle is the mechanism by which the human person emerges out of its tradition, and a new freedom is established for that which has emerged.[3]

III. Emergent Causation and Linguisticality

Truth and Method's account of what I have suggested can be identified as emergence is further clarified by a distinction between *world* and *environment*. For Gadamer, to have a *world* means to have an orientation toward it, or, in other words, to be able to establish

3. This interpretation of Gadamer's hermeneutics in *Truth and Method* supports Paul Ricoeur's attempt, in response to the debates between Gadamer and Habermas, to develop an account of hermeneutics that is compatible with a critique of tradition and authority (Piercey 2004, 263). I suggest, along with Ricoeur, that Habermas' distaste for tradition rests on a misunderstanding about the primordiality of hermeneutics, and his critique of ideology itself cannot be "detached from hermeneutic presuppositions" (Ricoeur 1991, 271). The interpretation I have provided of Gadamer's *Truth and Method* diffuses Habermas' concern that hermeneutics leaves no room for a critique of authority and tradition, suggesting that Gadamer's conception of the human person as emergent rests on the ontological possibility of creative critique and appropriation of an inherited tradition.

a critical stance with regard to it (2004 [1989], 440-441). In contrast, an *environment* in this context denotes the nexus of relations that exert causal pressure within which an organism finds itself (Ibid., 441). Gadamer explains that the freedom from this pressure is characteristically human and is effected by language: "To rise above the pressure of what impinges on us from the world means to have a language and to have a 'world'" (Ibid., 441). Whereas most animals experience a straightforward embeddedness in the environment, according to Gadamer, language allows its users a certain distance from particular aspects of this embeddedness, which affords the language user freedom with respect to the environment that simply embedded organisms cannot experience (Vilhauer 2010, 143).

Essentially, Gadamer argues a point here that is remarkably similar to one made by emergentist philosopher Warren Brown. Brown uses the idea of "action loops" to describe the way in which the behaviour of organisms never actually begins or ends; rather, it is a feedback loop in which an organism continually modulates its behaviour. The idea of action loops effectively reframes the discussion of causation, in that causation becomes modulation of pre-existing behaviour rather than the "triggering of action in an otherwise inert organism" (Brown 2007, 208). Significantly, Brown explains a basic structure of behaviour in terms of action loops in a way compatible with Gadamer's description of animals embedded in the environment. Further, Brown suggests that in more complex organisms (such as humans), who enjoy higher-level emergent properties (e.g., mind/conscious thought), there emerges multiple levels of supervisory systems that regulate and contain the more simple action loops (2007, 211). As a result, Brown considers humans to be able to rise above their simple action loops in the same way that Gadamer considers them capable of rising above the environment.

Converging on remarkably similar conclusions as Gadamer, Brown goes on to describe the ways in which language influences the emergence of supervisory systems. According to Brown, "scaffolding" refers to the ways in which organisms use their being-in-the-world to supplement mental processing. In an emergentist view of the mind, scaffolding suggests that mental activity in general is not simply internal to an organism but is fundamentally relational because it is supplemented to a certain extent by the environment. Language, suggests Brown, is "the primary form of external scaffolding of higher human mental abilities" (2007, 214). As such, language allows an organism to employ it as a tool with which to solve complex problems and to innovate in the world.

Gadamer continues on what seems to be an even more extreme path, claiming that "man's being-in-the-world is primordially linguistic" (2004 [1989], 440). However, understood in terms of the distinction between world and environment, this begins to

sound less extreme and more plausible. While being embedded in an environment may not necessarily be a linguistic phenomenon, rising above it (or developing supervisory systems for action loops) occurs as a result of the linguisticality of our being-in-the-world. Jean Grondin elucidates Gadamer's claim in this way: "putting into language is parallel to putting into understanding" (2003, 128). In other words, anything that is intelligible and understandable *can be described* and has significance. Grondin goes on to explain that for Gadamer, "everything presents itself to us under an aspect, because it concerns us and we participate in its manifestation" (2003, 149). When a human person is engaged in interpretation of something in the world, the object of interpretation presents itself "as" something to the interpreter (e.g., a cup *as* something to drink out of). Language is what allows the "as" structure of interpretation to bring objects in the world into relevance for an interpreter. Describing this structure of interpretation, Vilhauer explains that "[e]xplicit language is what allows some subject matter to be brought to presentation 'as' something, so that it becomes a distinct, meaningful part of our world" (2010, 143). Apart from language, creativity with respect to interpreting the world would not be possible, and we would be simply embedded in the environment rather than able to stand at a critical distance from certain aspects of it.

The primordially linguistic being-in-the-world of the human being is what creates the possibility for novel interaction with the world in general. Gadamer explains that when interpretation occurs and new understandings emerge, the interpreter is presented with "various possibilities for saying the same thing" (2004 [1989], 442). For example, learning that the world is round allows us to describe the world either according to our perceived experience or according to the new understanding (Ibid., 446). For Gadamer, however, description is not only wordplay; it is a manifestation of the ability to see new "as" structures, and therefore to envision new (equipotential) possibilities for activity in the world. Thus, as Robert Brandom suggests, because understanding, in the Heideggerian sense, is fundamental to being-in-the-world, and because language increases the possibilities for understanding, a new set of novel activities or performances is opened up by an interpreter's use of language (Brandom 1985, 186). Given the linguistic structure of human experience, we are able to draw conclusions about Gadamer's position on the origin of these novel performances and what mental or "downward" causation might look like for emergent Dasein.

Emergent Causation

To speak of mental causation is one way of describing the limits of human free will. Free will, understood as the ability for mental intentionality to effect action in the world, is necessarily a form of causation (Haag 2008, 113). Considering the human person to be emergent, strong emergentism allows for the development of free will out of a strong sense of downward causation. However, a strong emergentist also acknowledges the fundamental interdependence of the human person and the world. Describing the downward causal efficacy of cognition, emergentist philosopher Michael Silberstein argues that "[t]he social embeddedness of human cognition means that social features of an individual's life will help determine some of his or her psychological and neurochemical properties, not just the other way around" (2006, 213). That is, downward causation holds that the behaviour of an organism affects its lower-level functions at the same time that its lower-level functions affect its behaviour.

A weak emergentist position that advocates for a strong sense of global supervenience (that is, "the principle that two worlds that are microphysically identical will be or must be identical in all other macroscopic respects" [Silberstein 2006, 205]) also amounts to a denial of the possibility of *equipotentiality*. Equipotentiality is a term borrowed from Michael Polanyi that describes how a situation, a particular configuration of components in the world, may have more than one predetermined course, that is, that "lower level particulars can be regulated in more than one way, all of which have equal potential for producing a higher level performance (Dias 2008, 207). Openness to the possibility of equipotentiality, on the other hand, allows for moments of indeterminacy when an agent has no predetermined course of action and, therefore, has the freedom to choose from an array of equipotential possibilities. A strong emergentist position that emphasizes the embeddedness of cognition will likely resist global supervenience in favor of the possibility of some form of equipotentiality. The acceptance of strong emergence and downward causation opens up the possibility that an emergent may affect its originary system in an unexpected and unique way that may not be duplicated in an identical system (contrary to global supervenience). The denial of global supervenience as a blanket claim allows for the possibility that "under *exactly the same* circumstances agents are capable of doing different things" (Achim 2010, 187). Strong emergence, complete with downward causation, is the missing piece here that accounts for this ability.

I suggest that Gadamer's position amounts to a denial of global supervenience and an affirmation of equipotentiality, which arise out of his notion that the human person is capable of and is always engaged in "purposiveness," which, for Gadamer, is the ability to choose from a variety of suitable means to an end (2004 [1989], 470). Purposiveness

requires the ability to envision a variety of solutions to the same problem, or, in other words, it requires the ability to rise above the environment into a position of critical creativity. In this way, purposiveness is a special sort of downward causation, funded by the creativity enjoyed only by an emergent Dasein that is in a position of "*freedom from environment*" (Ibid., 442). Becoming (partially) dislodged from embeddedness in the pressures of the environment allows various possibilities to present themselves as equipotential, thereby opening up the possibility of the downward (mental) causation of purposiveness.

Conclusion: The Hermeneutics of Interdependence

To sum up, I have explored the constructive juxtaposition of strong emergence and Gadamer's hermeneutic ontology. Using Gadamer's notions of play, the hermeneutic circle, and linguisticality as touchstones, I hope to have demonstrated how each of these corresponds to a helpful way forward in defense of strong emergence. The benefits of a strong emergentist position include the affirmation of a robust sense of human free will and responsibility, along with a sense of human dependence on and belonging to the world and a tradition. Gadamer's compatibility with strong emergentism, like Heidegger's, offers a novel third way between dualism and reductive materialism, one that paves the way for an ontologically robust sense of ethical responsibility and indebtedness to the world and to the historical situation in which we find ourselves. This analysis of Gadamer's hermeneutic ontology in *Truth and Method* yields surprising evidence for Lauren Swayne Barthold's notion that, at bottom, Gadamer's hermeneutics is really about ethics:

> Understanding is ethical, then, to the extent to which it requires dialogical engagement with another; it is dialectical to the extent that we are caught in-between our own finitude and our longing to transcend it. Gadamer's dialectical hermeneutics helps us acknowledge our long forgotten kinship as the very offspring of Hermes. (Barthold 2010, 127)

As emergent persons, we are fundamentally constituted by our interactions with others and with the world, and yet we are inescapably responsible for our actions. Even more remarkable, however, is the fact that just as we effect change in our world, our world effects change in us.

References

Baetu, Tudor M. 2012. "Emergence, therefore antireductionism? A critique of emergent antireductionism." *Biology & Philosophy* 27 (3): 433–448.

Bedau, Mark A. 1986. "Cartesian Interaction." *Midwest Studies in Philosophy* 10 (1): 483–502.

Barthold, Lauren Swayne. 2010. *Gadamer's Dialectical Hermeneutics*. Lanham, MD: Lexington Books.

Brandom, Robert. 1985. "Freedom and Constraint by Norms." In *Hermeneutics and Praxis*, ed. Robert Hollinger, 173–191. Notre Dame, Indiana: University of Notre Dame Press.

Brown, Warren S. 2007. "The Emergence of Causally Efficacious Mental Function." In *Evolution and Emergence: Systems, Organisms, Persons*, ed. Nancey C. Murphy and William R. Stoeger, 198–226. Oxford: Oxford University Press.

Chalmers, David. 2006. "Strong and Weak Emergence." In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and P.C.W. Davies, 244–254. Oxford: Oxford University Press.

Clayton, Philip. 2006. "Conceptual Foundations of Emergence Theory." In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and P.C.W. Davies, 1–31. Oxford: Oxford University Press.

Deacon, Terrence W. 2007. "Three Levels of Emergent Phenomena." In *Evolution and Emergence: Systems, Organisms, Persons*, ed. Nancey C. Murphy and William R. Stoeger, 88–110. Oxford: Oxford University Press.

Dias, W. P. S. 2008. "Philosophical underpinning for systems thinking." *Interdisciplinary Science Reviews* 33 (3): 202–213.

Di Cesare, Donatella. 2013. *Gadamer: A Philosophical Portrait*. Translated by Niall Keane. Indianapolis, IN: Indiana University Press.

Di Francesco, Michele. 2010. "Two Varieties of Causal Emergentism." In *Emergence in Science and Philosophy*, ed. Antonella Corradini and Timothy O'Connor, 64–77. New York: Routledge.

El-Hani, Charbel Niño and Pihlström, Sami. 2002. "Emergence Theories and Pragmatic Realism." *Essays in Philosophy* 3 (2): 1–40.

Gadamer, Hans-Georg. 2004 [1989]. *Truth and Method*, 2nd rev. ed. Translated and revised by Joel Weinsheimer and Donald G. Marshall. New York: Continuum.

Georgiou, I. 2003. "The Idea of Emergent Property." *The Journal of the Operational Research Society* 54 (3): 239–247.

Gladwell, Malcolm. 2008. *Outliers: The Story of Success*. New York: Little, Brown and Company.

Grondin, Jean. 2003. *The Philosophy of Gadamer.* Trans. Kathryn Plant. Chesham England: Acumen Publishing Ltd.

Haag, James W. 2008. *Emergent Freedom: Naturalizing Free Will*. Göttingen: Vandenhoeck & Ruprecht.

Heidegger, Martin. 1996. *Being and Time*. Trans. Joan Stambaugh. Foreword by Dennis Schmidt. Revised. Albany: State University of New York Press.

Kauffman, Stuart A. 2008. *Reinventing the Sacred: A New View of Science, Reason and Religion*. New York: Basic Books.

Kim, Jaegwon. 2006. "Being Realistic about Emergence." In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and P.C.W. Davies, 189–202. Oxford: Oxford University Press.

Marino, Stefano. 2011. *Gadamer and the Limits of the Modern Techno-Scientific Civilization*. Bern: P. Lang.

Morgan, C. Lloyd. 1927 [1923]. *Emergent Evolution: The Gifford Lectures: Delivered in the University of St. Andrews in the Year 1922*. London: Williams and Norgate.

Murphy, Nancey. 2006. "Emergence and Mental Causation." In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and P.C.W. Davies, 227–243. Oxford: Oxford University Press.

O'Connor, Timothy and Jonathan D. Jacobs. 2003. "Emergent Individuals." *The Philosophical Quarterly* 53 (213): 540–555.

O'Connor, Timothy. 1994. "Emergent Properties." *American Philosophical Quarterly* 31 (2): 91–104.

Piercey, Robert. 2004. "Ricoeur's Account of the Habermas-Gadamer Debate." *Human Studies* 27 (3): 259–280.

Prosser, Simon. 2012. "Emergent Causation." *Philosophical Studies* 159 (1): 21–39.

Ramberg, Bjørn Torgrim. 1997. "The Source of the Subjective." In *The Philosophy of Hans-Georg Gadamer*, ed. Lewis Edwin Hahn, 459–471. Chicago, IL: Open Court.

Ricoeur, Paul. 1991. *From Text to Action*. Translated by Kathleen Blamey and John Thompson. Evanston: Northwestern University Press.

Silberstein, Michael. 2006. "In Defence of Ontological Emergence and Mental Causation." In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and P.C.W. Davies, 203–226. Oxford: Oxford University Press.

Spencer-Smith, Richard. 1995. "Reductionism and Emergent Properties." *Proceedings of the Aristotelian Society* 95:113–129.

Stephan, Achim. 2010. "Are Deliberations and Decisions Emergent, if Free?" In *Emergence in Science and Philosophy*, ed. Antonella Corradini and Timothy O'Connor, 180–189. New York: Routledge.

Sterelny, Kim. 2001. *Dawkins vs. Gould: Survival of the Fittest*. Cambridge, MA: Totem Books.

Van Gulick, Robert. 2007. "Who's In Charge Here? And Who's Doing All the Work?" In *Evolution and Emergence: Systems, Organisms, Persons*, ed. Nancey C. Murphy and William R. Stoeger, 74–88. Oxford: Oxford University Press.

Vilhauer, Monica. 2010. *Gadamer's Ethics of Play: Hermeneutics and the Other*. Lanham, MD: Lexington Books.

Journal of Cognition and Neuroethics

Managing Serious Incidental Findings in Brain-Imaging Research: When Consent for Disclosure is Declined

Chiji Ogbuka
University of Maryland University College

Biography

Chiji Ogbuka is the Project Manager of Regulatory Affairs and Medical Writing at Health Decisions Inc., Durham NC. Chiji is also an Adjunct Associate Professor of Regulatory Affairs at University of Maryland University College, Adelphi, MD, and is currently a doctoral candidate at St. Louis University, St. Louis, MO, pursuing doctoral studies in Health Care Ethics. His research and professional interests are in Neuroethics, Medical Writing, Pharmaceutical Drug Development, and Regulatory Affairs.

Publication Details

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

Citation

Ogbuka, Chiji. 2014. "Managing Serious Incidental Findings in Brain-Imaging Research: When Consent for Disclosure is Declined." *Journal of Cognition and Neuroethics* 2 (2): 51–59.

Managing Serious Incidental Findings in Brain-Imaging Research: When Consent for Disclosure is Declined

Chiji Ogbuka

Abstract

This paper focuses on the management of *serious* neuro-imaging incidental findings (NIFs) when a participant declines consent. To prevent severe neurological complications, serious NIFs necessitate immediate clinical referral. When consent for disclosure is explicitly declined, researchers face a significant dilemma in assessing ethical obligations of beneficence relative to the participant's autonomous choice. Relying on the Belmont principles, I adapt Henry Richardson's theory of *specification* to argue that the researcher's duty of beneficence is shaped by the expressed autonomous choice of the participant. To best avoid such a conflict of principles and norms, researchers should specifically address consent for disclosing serious NIFs as a criterion for study participation.

Keywords

Neuro-imaging incidental findings (NIFs), serious, brain-imaging research, Richardson, informed consent, autonomy, beneficence

Introduction

Brain-imaging research has benefited tremendously from innovations in both functional and structural neuro-imaging technologies. Such technologies have augmented the potential for high-resolution imaging and mapping of intracranial surface structures, brain substructures, and neural correlates with clinical and anatomic precision. The application of neuro-imaging technologies in research has lead to the discovery of novel therapies for treating neurologic and psychiatric diseases. Together with these achievements, scientific and ethical challenges have become evident in the application of neuro-imaging technologies particularly within research. One of such challenges is the ethical dilemma surrounding a researcher's obligation to disclose [or not disclose] a *serious* neuro-imaging incidental finding (NIF) when a research participant expressly declines consent for disclosure.

To resolve this conflict of ethical principles and the norms that derive from them, I adopt the ethical tool of *specification*—a theory of practical reasoning developed by Henry Richardson (1990). I argue that when consent for disclosure of NIFs is declined by

a competent participant, the obligations of beneficence is shaped and specified by the expressed autonomous desire of the participant. The real possibility of detecting serious health-related incidental findings in brain-imaging research engenders a need for careful planning and preparation in the design of brain-imaging studies. To avoid the dilemma of conflicting ethical norms in this context, I propose that researchers specifically address consent for disclosure of clinically urgent NIFs as a criterion for study participation.

Neuro-Imaging Incidental Findings (NIFs)

An NIF is defined as a health-related discovery or anomaly in the neuro-imaging scan of a research participant that is not directly relevant to the variables investigated in the research study. NIFs are often identified as unexpected anomalies on a brain scan such as an inflammatory lesion, a vascular malformation, a neoplasm, intracranial aneurysm, cyst, or a host of other potentially symptomatic cerebro-vascular disease-markers (Morris et al. 2009). It is *serious* when it is clinically significant, analytically valid, poses an immediate health or pathological risk of danger, and is actionable. The human brain is such that the existence of anomalous electrical, structural or biochemical variations could indicate adverse conditions like memory loss, paralysis, seizures, neuromuscular diseases, or other potentially serious neurological disorders. As such, detecting, disclosing, and properly managing such anomalies are critical and of great ethical concern within neuro-imaging research. While incidental findings (IFs) in general have been conventionally construed as *incidental* to research, limitations of such descriptions are unquestionable (Parker, 2008). Depicting such findings as *incidental* generates practical challenges about whose health is at stake and whose interests deserve priority (Illes and Chin 2008).

The importance of anticipating and managing serious NIFs in research has been highlighted in empirical studies, government reports, and institutional research guidelines. Proper planning, adequate professional expertise, communication, improved consent practices, and transparency within neuro-imaging research are critical (National Institutes of Health and Stanford University 2005). Likewise, empirical studies underscore the necessity for urgent clinical referral when a serious NIF is confirmed. In one study of 1000 asymptomatic volunteers (between ages 3-83) from a variety of NIH research protocols, 180 cases of NIFs were reported, 18 of which required routine referral, and 11 required urgent referral due to tumors and lesions (Katzman, Dagher, and Patronas 1999). Similarly, a retrospective review of 151 MRI studies on healthy volunteers from previous studies indicated a 6.6% NIF incidence rate requiring clinical referral with 3 cases of clinical urgency (Illes et al. 2004). In addition, Yue and colleagues (1997) reviewed

3672 image scans in a population-based study of asymptomatic elderly individuals and reported 64 cases of NIF with only 9 *serious* anomalies requiring urgent surgical referral.

Despite their empirical and clinical significance, there is divergence about managing *serious* NIFs. The high incidence of false positives, the possibility of triggering burdensome or costly interventions, and the potential for ambiguous findings can complicate disclosure (Royal and Peterson 2008). Consequently, the obligations neuro-imaging researchers owe participants given the overall aims of *generalizable knowledge* intrinsic to research need to be further specified. Some authors advance a fiduciary relationship requiring certain clinical care and equipoise standards (Weijer and Miller 2003). Others propose an ancillary care framework grounded in partial entrustment (Rangel 2010; Richardson and Belsky 2004). One position underscores the researcher's obligations as a responsibility with binding professional implications (Miller, Mello, and Joffe 2008). These frameworks, though valuable, only apply when a participant consents to disclosure or when the consent process fails to address serious NIFs. However, when consent is expressly declined, a different kind of ethical assessment is necessitated to *specifically* address the management of *serious* NIFs during informed consent.

Specification

Richardson's (1990) notion of *specification* involves a systematic method of practical reasoning from abstract norms to concrete actionable guides by constantly shaping and substantiating the applicable norms with content. Specification presupposes the existence of a set of ethical norms; it then proceeds to determine how these norms apply in shaping action, particularly when these norms conflict. With respect to research, Richardson proposes a *protean research limiting principle* and examines its ramifications from a less restrictive and more restrictive perspective. The protean principle states that "it is impermissible to engage in research on human subjects unless the principles of autonomy, beneficence, and justice are adequately satisfied" (301). On this principle, we fruitfully can recast the debate on how both autonomy and beneficence can be specified. Specification articulates the different interpretive options visible in the two ramifications of the protean principle's notion of adequate satisfaction: "it is impermissible to engage in research on human subjects unless the principles of autonomy, beneficence, and justice are *satisfied on balance* (less restrictive); it is impermissible to engage in research on human subjects unless we do so in a way that *respects their autonomy, proceeds justly, does no (intentional harm), and produces (significant) benefits* (more restrictive)" (301).

Practical Considerations

Human subjects' research is guided by ethical principles (autonomy, beneficence, and justice) outlined in the Belmont Report which provide a basis for assessing obligations. These principles serve as heuristics that offer action-guiding content when specified in the form of norms (Meslin, Sutherland, Lavery, and Till 1995). Since practical reasoning involves a means-end assessment of action-guides in particular contexts, specifying one's end helps to focus the process of attaining that end. Within research, the obligation of beneficence is shaped and specified by the expressed autonomous choice of the participant. Informed consent formally establishes the relationship between the neuro-imaging researcher and participant. Consent is necessary for disclosing serious NIFs. As such, any action to manage a serious NIF should involve a re-assessment of a participant's denial of consent in light of the discovered "incidental" abnormality.

The first step in applying specification involves deliberation on the morally relevant facts pertaining to beneficence and autonomy. They include:

a. the discovery of a serious NIF;
b. the denial of consent for disclosure;
c. the possibility of immediate neurological harm and
d. the responsibility of the researcher/institution in light of the discovery.

From the protean principle, two norms emerge, one less restrictive, the other, more restrictive.

a. less restrictive specification – it is impermissible to disclose serious NIFs to participants unless the requirements of consent are adequately satisfied;
b. more restrictive specification – it is impermissible to disclose serious NIFs to participants unless we do so in a way that respects their autonomous choice and maximizes benefits.

The presumption from beneficence is to maximize benefits and minimize risks. When consent for disclosure is expressly denied, the obligation of beneficence becomes even more difficult. Beneficence does not simply override the duty to respect a participant's autonomous choice. Moreover, beneficence cannot be coercive with the objective of maximizing the health benefit of the participant. A denial of consent practically limits the scope of beneficence. Since serious NIFs are detected in the form of unexpected anomalies—lesions, aneurysms, vascular defects, etc., managing them can be significant. The human brain is such that the existence of anomalous electrical, structural, and/or biochemical variations could be indicative of critical neurological conditions needing immediate referral. As a result, overlooking such anomalies (doing nothing) cannot be adequate. The nature and integrity of brain-imaging research, demands a higher ethical

standard: using a brain scan from a supposedly healthy volunteer discovered to have a tumor or intracranial malformation is, at the very least, problematic (Illes and Chin 2008). As a matter of fact, discovering a serious NIF shifts the priority from research participation to an urgent need for clinical care, at least for the given participant. While it is impermissible to foist disclosure of serious NIFs on participants, doing nothing is ethically untenable. A refusal of disclosure during initial consent is inadequate to satisfy the requirements for managing significant and serious NIFs. Though re-consenting in light of an actual (not potential or statistic) finding may be practically and logistically difficult, it is necessary to pursue immediate clinical referral. A refusal of consent at this point severely jeopardizes the participant's continued participation in the study.

Implications for Consent

Addressing the disclosure and management strategy for potential NIFs during informed consent must be an essential provision and requirement for conducting brain-imaging research. While the nature, incidence, sensitivity, and severity of NIFs may vary across brain-imaging research settings, their clinical significance, utility, and actionability tend to fall within three major categories of classification requiring immediate/urgent referral, routine referral, and/or no referral. The meanings and implications of these categories should be explained and subjects should provide individual consent/non-consent (perhaps in the form of checkboxes or initials) to be informed about NIFs that require an action plan for each of these categories. One reason for this is that, general consent to be informed about NIFs may not specifically address clinical considerations of action plan that require urgent, routine, or no referral.

It is important to note that the potential for discovering serious NIFs in brain-imaging research demands a careful assessment of suitability for study participation based on participants' disposition for consent. If the consent process adequately educates participants on the nature, empirical incidence, and significance of NIFs, and clearly specifies areas of individual consent/non-consent for disclosure of such findings, a legitimate refusal of consent for disclosure, should be grounds for study exclusion. Obviously, this position is practically difficult given concerns about scientific validity and social value. The nature and integrity of neuro-imaging research nonetheless calls for a rigorous consent process. Likewise, the possibility of a serious adverse event looms large if a serious NIF is not managed immediately.

Participants as such, should be given an option on the consent form to voluntarily opt out of participating in the study, if they do not wish to consent to disclosure for

serious NIFs or be contacted in the future, if a serious NIF is found. This determination is clearly distinct from categorically excluding participants from the study based on a perceived potential for a serious NIF. This latter situation should be discouraged since it violates the ethical principle justice which operationalizes fairness and equality in the selection of participants and the distribution of burdens.

Conclusion & Recommendations

At minimum, a serious NIF detected in brain-imaging research deserves medical attention. It ultimately changes the context of research for the given participant. The informed consent process should explicitly and adequately address the incidence/potential for serious NIFs, invite constructive discussions from participants, discuss the real possibility of re-contacting participants, and map out any clinical follow-up plan for serious NIFs. In sum, a proactive and preparatory strategy for managing NIFs is certainly preferred to a reactive one.

I recommend that consent documents address different consent levels for NIFs (using initials or checkboxes, for example) with varying degrees of clinical referral need:

- For NIFs that require no clinical referral (have low indication of risk), consent for disclosure may not be required;
- For NIFs that require routine referral, requirements for consent for disclosure should be based on individual assessments of severity by a competent clinician;
- For NIFs that require urgent and immediate referral (high risk of harm), consent must be required for participation in the study. If this box is not initialed or checked, participation in the study should be prevented.

References

Illes, Judy, and Vivian Nora Chin. 2008. "Bridging philosophical and practical implications of incidental findings in brain research." *Journal of Law, Medicine, & Ethics* 36 (2): 298–304.

Illes, Judy, Rosen, A., Huang, L., Goldstein, R., Raffin, T., Swan, G., & Atlas, S. 2004. "Ethical consideration of incidental findings on adult brain MRI in research." *Neurology* 62 (6), 888–890.

Katzman, Gregory, Azar Dagher, and Nicholas Patronas. 1999. "Incidental findings on brain magnetic resonance imaging from 1000 asymptomatic volunteers." *Journal of American Medical Association* 282: 36–39.

Meslin, Eric, Heather Sutherland, James Lavery, and James Till. 1995. "Principlism and the ethical appraisal of clinical trials." *Bioethics* 9 (5): 399–418.

Miller, Franklin, Michelle Mello, and Steven Joffe. 2008. "Incidental findings in human subjects research: What do investigators owe participants?" *Journal of Law, Medicine, & Ethics* 36 (2): 271–279.

Morris, Zoe, William Whiteley, William Longstreth, Frank Weber, Yi-Chung Lee, Yoshito Tsushima, Hannah Alphs, Sussane Ladd, Charles Warlow, Joanna Wardlaw, and Rustam Al-Shahi Salman. 2009. "Incidental findings on brain magnetic resonance imaging: Systematic review and meta-analysis." *British Medical Journal* 339: b3016.

National Institutes of Health & Stanford University. 2005. Proceedings from conference: "*Detection and Disclosure of Incidental Findings in Neuroimaging Research.*" Bethesda, MD.

Parker, Lisa. 2008. "The future of incidental findings: Should they be viewed as benefits?" *Journal of Law and Medical Ethics* 36 (2): 341–351.

Rangel, Erica. 2010. "The management of incidental findings in neuroimaging research: Framework and recommendations." *Journal of Law and Medical Ethics* 38 (1): 117–26.

Richardson, Henry. 1990. "Specifying norms as a way to resolve concrete ethical problems." *Philosophy & Public Affairs* 19 (4): 279–310.

Richardson, Henry, and Leah Belsky. 2004. "The ancillary care responsibilities of medical researchers: An ethical framework for thinking about the clinical care that researchers owe their subjects." *Hastings Center Report* 34 (1): 25–33.

Royal, Jason, and Bradley Peterson. 2008. "The risks and benefits of searching for incidental findings in MRI research scans." *Journal of Law & Medical Ethics* 36 (2): 305–314.

Weijer, Charles, and Paul Miller. 2003. "Therapeutic obligation in clinical research." *The Hastings Center Report* 33 (3): 3–4.

Yue, N., Longstreth, W., Elster, A., Jungreis, C., O'Leary, D., & Poirier, V. (1997). "Clinical serious abnormalities found incidentally at MR imaging of the brain: Data from the Cardiovascular Health Study." *Radiology* 202: 41–46.

Journal of Cognition and Neuroethics

Is Neuroscience Relevant to Our Moral Responsibility Practices?

Joseph M. Vukov
Fordham University

Biography
Joe Vukov is a PhD candidate in the philosophy department at Fordham University. His research interests lie primarily in philosophy of mind, metaphysics, and bioethics.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). November, 2014. Volume 2, Issue 2.

Citation
Vukov, Joseph M. 2014. "Is Neuroscience Relevant to Our Moral Responsibility Practices?" *Journal of Cognition and Neuroethics* 2 (2): 61–82.

Is Neuroscience Relevant to Our Moral Responsibility Practices?

Joseph M. Vukov

Abstract

Some psychologists and philosophers have argued that neuroscience is importantly relevant to our moral responsibility practices, especially to our practices of praise and blame. For consider: on an unprecedented scale, contemporary neuroscience presents us with a mechanistic account of human action. Furthermore, influential studies – most notoriously, Libet et al. (1983) – seem to show that the brain decides to do things (so to speak) before we consciously make a decision. In light of these findings, then – or so some have argued – we ought to revise our practices of praise and blame. In the current paper, I argue that this conclusion is unwarranted. The reason is that the argument for it depends on controversial non-empirical premises, premises we need not accept. I suggest, however, that neuroscience does bear on our moral responsibility practices in one important, if less revisionary, way. In particular, neuroscience offers a new kind of evidence for determining when agents should be held exempt from our normal moral responsibility practices.

Keywords

Moral responsibility; neuroscience; incompatibilism; Libet; conscious control

Is Neuroscience Relevant to Our Moral Responsibility Practices?

Some psychologists and philosophers have argued that neuroscience is importantly relevant to our moral responsibility practices, especially to our practices of praise and blame. Specifically, they have argued for the following theses:

i. On an unprecedented scale, contemporary neuroscience presents us with a mechanistic account of human action. But if our actions can be accounted for mechanistically, this gives us reason to revise the ways in which we hold agents morally responsible.

ii. Influential studies – most notoriously Libet et al. (1983) – show that the brain decides to do things (so to speak) before we consciously make a decision. But since conscious control is necessary for holding agents morally accountable, these studies suggest the need for revising our moral responsibility practices.

iii. Still other studies have discovered neural correlates for specific psychological abnormalities. Neuroscience thus offers a genuinely new kind of evidence for

deciding when particular individuals meet criteria for exemption from our moral responsibility practices.

The worries presented in (i-iii) therefore make it seem that neuroscience is importantly relevant to our moral responsibility practices. I will argue, however, that neuroscience is in fact relevant these practices only in the way spelled out in (iii). More specifically: I will argue (in Section 1) that (i) is an important consideration only if we assume a number of controversial metaphysical premises. And since metaphysical premises cannot be supported empirically, neuroscience cannot by itself raise the worry presented by (i). The worry presented by (ii), in turn, is a concern only if we take a lack of conscious control to undermine our moral responsibility practices. But, or so I will argue (in Section 2), a lack of conscious control doesn't undermine our moral responsibility practices. This leaves (iii), then, as the only way neuroscience bears on our moral responsibility practices (Section 3). And while (iii) is not as revisionary as either (i) or (ii), recognizing this specific role neuroscience can play still has important implications.

1. Neuro-Revisionism

1.1 The Argument for Neuro-Revisionism

A number of authors have argued that the mechanistic picture of human action emerging from contemporary neuroscience should lead us to revise our moral responsibility practices (Farah & Heberlein 2006; Wright 1994; Greene & Cohen 2004). Let's call anyone who argues for this conclusion a *Neuro-Revisionist*. Why should we accept Neuro-Revisionism? Greene & Cohen (2004) articulate a standard version of the argument – it can be reconstructed as follows:

P1. Our actions can be accounted for mechanistically.

P2. If our actions can be accounted for mechanistically, then there is no free will.

C1. Therefore, there is no free will.

P3. If there is no free will, our moral responsibility practices should be revised.

C2. Therefore, our moral responsibility practices should be revised.

Let's consider why we might be tempted to accept each premise:

P1, first of all, claims that our actions can be accounted for *mechanistically*. I'll follow Bechtel and Richardson's (1993) definition of a 'mechanistic explanation' as any explanation that "propose[s] to account for the behavior of a system in terms of the functions performed by its parts and the interactions between these parts" (Bechtel and Richardson 1993, 17). And indeed, new work in neuroscience seems to give just such an explanation of many of our actions. Consider, for instance, the story we are now able to tell about the simple action of reading aloud a written word:

> Light reflects off the page and strikes the retina. Different retinal cells react to various wavelengths of light, which in turn leads to different patterns of firing in the ganglion cells. These cells then innervate the lateral geniculate nucleus of the hypothalamus, which in turn sends a tract to the primary visual cortex, V1. V1 then innervates the angular gyrus, which decodes the visual information and transmits this information to Broca's area via the arcuate fasciculus. Broca's area, in turn, creates a motor plan for speaking the word and transmits this plan to the primary motor cortex, where the motor plan is implemented. (Breedlove et al. 2010)

This story, it seems, thus accounts for the action exclusively in terms of neurological function; it accounts for the action *mechanistically*. And while were are not yet able to tell a story to this degree of detail about complex human actions – actions such as falling in love, composing a poem, or playing a game of chess – advances in neuroscience suggest that such a description is possible, if not inevitable.

Turn next to P2. Greene & Cohen defend this premise by appeal to an intuitive understanding of free will (which is, as we will see, problematic – but more on this later). Specifically, they argue that "people regard actions only as fully free when those actions are seen as robust against determination by external forces" (Greene & Cohen 2004, 1780). My actions are intuitively free, in other words, only when they have their source in *me and me alone*. On Greene & Cohen's account, then, actions performed with intuitive free will are importantly different from actions that can be accounted for mechanistically. When we account for an action mechanistically, after all, we account for it not in terms of an agent's volition, but rather in terms of the operations of sub-personal mechanisms. If my actions can be accounted for mechanistically, then, they could not have their source in *me* – they could not be intuitively free. Equivalently then, if our actions can be accounted for mechanistically, there are not performed freely.

Last, consider P3. Greene & Cohen defend this premise by observing that our current moral responsibility practices are largely founded on a presumption of free action. Consider, for instance, our current practices of punishment – these practices, Greene & Cohen argue, are founded on the assumption that "people…*deserve* to be punished, and that is why we punish them" (Greene & Cohen 2004, 1776). Intuitively, however, a person cannot *deserve* to be punished for an action that was not freely performed.[1] Insofar then as our current practices of punishment are founded on a presumption of free will, it follows that if there is no free will, we must revise these practices.[2]

And from this, Neuro-Revisionism follows – recent discoveries in neuroscience provide us with reason to revise our moral responsibility practices. If we accept P1-P3, we must therefore accept that "for ethics, the only alternative we can see is a shift to a more utilitarian approach [to our moral practices]" (Farah & Heberlein 2007, 46), that "advances in neuroscience are likely to change the way people think about human action and…responsibility" (Greene & Cohen 2004, 1784).

1.2 Response to the Argument for Neuro-Revisionism

Above, I outlined the argument for Neuro-Revisionism. In the current section, I hope to show that the argument's conclusion is unwarranted. The reason is that the argument depends on the truth of a number of controversial metaphysical premises, premises we need not accept.

To clear up some terminology as I will be using it: *determinism* claims that given the state of the world at any given time and the laws of nature, there is only one possible resultant state of the world at any given later time.[3] *Incompatibilism* is the thesis that free will is not compatible with determinism; incompatibilism, therefore, claims that if determinism is true, our actions cannot be free.[4] *Compatibilism*, on the other hand, claims that even if determinism is true and our actions are thus causally determined, it is still

1. For a similar and more thoroughly defended argument for this position, see Pereboom (2013).

2. Greene & Cohen make some obviously controversial claims in the course of defending P3. But as nothing they say here is relevant to my current purposes, I allow their claims to remain unchallenged.

3. This definition may be contended, of course. For my purposes, however, it is sufficient to have a working definition of determinism. And I take it that this definition – while not beyond reproach – is by no means unorthodox either.

4. For some influential incompatibilist accounts, see Chisholm (1964), van Inwagen (1975), and Kane (1996).

possible for these actions to be free.[5] Importantly, then, compatibilists and incompatibilists do not disagree about the truth or falsity of determinism; rather, the positions disagree about the status free will would have were determinism true.

With this terminology in hand, let's turn back to Greene & Cohen's argument for Neuro-Revisionism. Specifically, consider P2 of their argument – this premise claims that 'if our actions can be accounted for mechanistically, then there is no free will.' As we saw, Greene & Cohen defend this premise by appealing to an intuitive understanding of free will. Intuitively free actions, they claim, are importantly different from actions that can be accounted for mechanistically. At first blush, this reasoning may seem sound. In fact, however, it depends on two suppressed metaphysical premises.

The first of these premises forms the foundation for Greene & Cohen's claim that "people regard actions only as fully free when those actions are seen as robust against determination by external forces" (Greene & Cohen 2004, 1780). For this claim, as should now be apparent, is not innocent. Compatibilists, after all, *do* regard free actions as compatible with determinism – compatibilists can therefore argue that actions have their source in *me* even when those actions are determined.[6] In the course of arguing for P2, Greene & Cohen must therefore assume:

> **Incompatibilism**: If our actions are determined, then there is no free will.

Greene & Cohen, to their credit, acknowledge that their argument for P2 assumes **Incompatibilism**. They urge, however, that contemporary neuroscience reveals our *true* intuitions – they argue that the mechanistic picture of human action presented by neuroscience supports:

> ...a powerful moral intuition that…compatibilist philosophies sweep under the rug…[P]eople regard actions only as fully free when those actions are seen as robust against determination by external forces. But if determinism…is true, then no actions are truly fee because forces beyond our control are always sufficient to determine behavior. Thus,

5. For some influential compatibilist accounts, see Ayer (1954), Frankfurt (1971), Watson (1996), Strawson (1963), Wolf (1990), Fischer (1987; 1994), Fischer & Ravizza (1998).

6. Although they need not argue in this way – it is possible to be a compatibilist (and so affirm that our actions are free) and yet deny that we are the source of our actions.

> intuitive free will is [incompatibilist], not compatibilist. (Greene & Cohen 2004, 1780)[7]

According to Greene & Cohen, in other words, when we carefully consider the mechanistic picture painted by neuroscience, we intuitively see that free will finds no place in it.

This argument for **Incompatibilism** won't work, and for a number of reasons. *First*, Greene & Cohen's claim that our raw moral intuitions are decidedly incompatibilist seems to be false on purely empirical grounds, or – at the very least – more empirically complicated than they make it seem. A series of recent studies have found that untutored intuitions are in many cases *compatibilist*. Nahmias et al. (2005), for instance, found that people tend to judge agents morally responsible for actions even when these actions are construed as determined. Similarly, Nichols and Knobe (2007) found that people's intuitions tend to be compatibilist as long as moral scenarios are posed in a concrete, emotional manner. These findings, of course, are not beyond reproach – Nahmias et al. (2007), for instance, found that our intuitions *do* tend towards incompatibilist when scenarios are couched in mechanistic neural terms. This particular study therefore supports Greene & Cohen's hypothesis. Taken as a whole, however, recent empirical work suggests that our raw moral intuitions are much more nuanced that Greene & Cohen assume.

Suppose though that our intuitions *were* incompatibilist in just the way Greene & Cohen assume. Even then, this does not give us good reason to accept **Incompatibilism**. For while consistent incompatibilist intuitions may count as *prima facie* evidence in favor of **Incompatibilism**, intuitions – especially moral ones – are notoriously theory-laden; they often rest on unfounded assumptions, confusions, and equivocations. Most compatibilist arguments, furthermore, consist precisely in diagnosing the faulty reasoning lying at the heart of incompatibilist intuitions. So then, by treating these intuitions as decisive rather than *prima facie* evidence in favor of incompatibilism, Greene & Cohen ignore the standard compatibilist reply. And in ignoring this reply, they fail to offer an argument that any compatibilist would accept.

Now, let me be clear about what I am arguing here: I am not, in the first place, claiming that moral intuitions are irrelevant to our moral reasoning; clearly, they are. And I am also not suggesting that **Incompatibilism** is false; it could very well be that our incompatibilist intuitions are perfectly accurate. Rather, I am pointing out that these intuitions cannot – as Greene & Cohen suppose – be taken as decisive in the debate between compatibilists and incompatibilists; they should rather be an occasion for this

7. This claim is echoed in Kane (1999) and Strawson (1986).

debate. Insofar as the argument Greene & Cohen give for Neuro-Revisionism depends on **Incompatibilism**, it therefore depends on a metaphysical premise they do not adequately defend.

And things get worse. The reason is that **Incompatibilism** does not entail P2 – P2 claims that 'if our actions can be accounted for *mechanistically*, then there is no free will' while **Incompatibilism** claims that 'if our actions are *determined*, then there is no free will.' To move from **Incompatibilism** to P2, Greene & Cohen must therefore assume the following thesis:

> **Mechanism entails Determinism**: If our actions can be accounted for mechanistically, then our actions are determined.

But it is by no means obvious that **Mechanism entails Determinism** is true. The reason is simple: it is not obvious that just because we can account for an action mechanistically that the mechanisms in play are *themselves* deterministic. It is possible, for instance, that our actions could ultimately be explained in terms of the behavior of *indeterministic* mechanisms – and indeed, this is precisely the position taken by event-causal libertarians (Wiggins 1973; Ekstrom 2000; Kane 1996). In taking on **Mechanism entails Determinism**, Greene & Cohen therefore find themselves suppressing yet another controversial and inadequately defended metaphysical premise.

The upshot of this is that Greene & Cohen's argument does not establish Neuro-Revisionism – Neuro-Revisionism follows from their argument only if we accept both **Incompatibilism** and **Mechanism entails Determinism**. But determining the truth of *these* theses is a distinctively philosophical project – these theses, after all, are distinctively metaphysical, and there is no clear way in which neuroscience could be of use answering them. In the end, then, the argument for Neuro-Revisionism is a poor one, and we need not accept its conclusion. If Neuro-Revisionists are to succeed in showing that neuroscience is relevant to our moral responsibility practices, they must first address philosophical questions about free will, determinism, mechanism, and moral responsibility – the very questions philosophers have focused on from the beginning.

2.0 Libet et al. (1983), the Principle of Conscious Control, and Moral Responsibility Skepticism

2.1 The Argument from Libet for Moral Responsibility Skepticism

I argued in Section 1 that the mechanistic picture of action presented by neuroscience is not capable of leading us to revise the way in which we hold agents morally accountable. Or it is not capable of doing so, at any rate, without some hefty philosophical lifting. But there is another way to raise the worry that findings in neuroscience give us reason to revise our moral responsibility practices. Furthermore, this way of raising the worry does not depend on any metaphysically-contentious claims concerning the relationship between mechanistic explanation, determinism, and free will. It rather points to specific findings in neuroscience – most centrally, those of Libet et al. (1983) – and argues that these findings universally undermine moral responsibility. I'll call anyone who argues along these lines a *Moral Responsibility Skeptic*.

Moral Responsibility Skepticism lurks in a great deal of the literature surrounding Libet's findings (Libet 1983; Libet 2011; Banks & Isham 2011; Hallett 2011; Pockett 2004, Roediger et al. 2008; Spence 2009; Wegner 2002). As Bayne (2011) points out, however, full-fledged arguments for the position rarely boil to the surface – Moral Responsibility Skeptics rarely articulate precisely *how* Libet endangers moral responsibility. In this section, then, I will consider a generic version of the argument for Moral Responsibility Skepticism, one that – while generic – seems to lie at the base of the more particular worries raised by Libet and his followers. The argument proceeds as follows:

> P1: For any action, *a*, and any subject, S, S can be held morally responsible for *a* only if *a* is caused by the conscious decision of S.
>
> P2: There is no action, *a,* such that *a* is caused by the conscious decision of a subject.
>
> C1: Therefore, there is no action, *a*, such that there is some subject who can be held morally responsible for *a*.

Let's work through the premises. Call P1 the Principle of Conscious Control [PCC]; according to PCC, if an agent does not cause an action by his or her conscious control, that agent cannot be held morally responsible for the action. Velleman (1992) offers a reason for why we might adopt something like PCC. In the course of cashing out a scenario – one in which he finds himself yelling at a friend without having made a conscious decision to do so – Velleman reasons:

> …viewing the decision [to yell at my friend] as directly motivated by my desires, and my behaviors as directly governed by the [unconscious]

> decision…leads to the thought that as my words become more shrill, it was my resentment speaking, not I. (Velleman 1992, 465)

According to Velleman, in other words, events that result from unconscious decisions cannot be understood as genuine actions at all. This is because, on Velleman's view, insofar as a decision is made unconsciously, it is not genuinely *one's own*. But if an event is not an action at all, it is certainly not the sort of thing for which agent could be held responsible. Hence, if Velleman is right, we seem to have good reason to suppose that conscious control is a necessary condition for holding agents morally responsible.

Velleman's considerations aside, there is an intuitive plausibility to PCC. The paradigmatic actions for which we hold agents morally responsible, after all, are those which seem to be caused by the conscious decision of an agent; conversely, the paradigmatic actions for which we don't hold agents morally responsible are those which do not seem to be caused by the conscious decision of an agent. Consider: we hold agents morally responsible for lying, murdering, theft, donating to charity, and so on; we don't hold agents morally responsible for twitches, schizophrenic episodes, compelled misdeeds, and so on. And it seems that the latter actions differ from the former precisely in that the former are caused by the conscious decision of the agent in a way that the latter are not. More, of course, needs to be said about PCC. But the motivation behind adopting P1 should be clear.

On then to P2. This is the premise Libet et al. (1983) purports to confirm, so it's worth reviewing Libet's findings: in the electroencephalography [EEG] study, participants were instructed to relax and then – whenever they chose – to self-initiate "quick, abrupt flexion of the fingers and/or wrists of [the] right hand" (Libet et al 1983, 625). Participants were also asked to report the moment at which they first felt the desire to perform this movement. The study found that "with few exceptions, onset of [cerebral activity] occurred before reported awareness time by substantial amounts of time" (Libet et al. 1983, 634). Libet et al. (1983) thus concludes that:

> …the brain evidently 'decides' to initiate or, at the least, prepare to initiate the act at a time before there is any reportable subjective awareness that such a decision has taken place. It is concluded that the cerebral initiation of even a spontaneous voluntary act, of the kind studied her, can and usually does begin unconsciously. (640)

Subsequent studies have echoed Libet's findings (Soon et al. 2008; Haynes 2011; Haggard & Eimer 1999; Keller & Heckhausen 1990; Lau et al. 2004). Soon et al. (2008) is especially

noteworthy: in this functional magnetic resonance imaging [fMRI] study, participants' neural activity was monitored during simple motor decisions. The study found neural activations predictive of the decision up to *10 seconds* before the decision was reported as having been made consciously. Soon et al. (2008) thus conclude, following Libet, that "a network of high-level control areas can begin to shape an upcoming decision long before it enters awareness" (543). Taken together, these studies thus suggest that in simple motor actions, a decision can *seem* to be caused by a conscious decision even while it is in fact caused by some earlier neural activation. P2 then generalizes these findings to include all actions.

And this is all the Moral Responsibility Skeptic needs. For from P1-P2, C1 follows straightforwardly: no one can be held morally responsible for anything.

2.2 Against Moral Responsibility Skepticism

I want to resist the argument for Moral Responsibility Skepticism. There are two general strategies for doing so. First, one could maintain that P2 is false. Adopting this strategy thus means claiming that there are some actions that are directly caused by the conscious decisions of agents. This position remains viable, even in the face of contemporary neuroscience. For Libet's findings simply do not confirm the totalizing claim made by P2, the claim, that is, that there are *no* actions that are caused by conscious decisions. At most, Libet's findings confirm that there are some, relatively simple motor decisions that are not caused by conscious decisions. And the findings may not even show this – Libet's study has been challenged on an empirical level, and these contrary findings need to be taken into account before adopting anything like P2.[8]

Still, the strategy of rejecting P2 outright has its weaknesses. In the first place, while Libet and his followers draw inarguably hasty conclusions, they aren't wrongheaded in worrying that the study has serious implications. Here's why: I take it that our best evidence for not-P2 is that it *seems to us* as if our actions are often caused by conscious decision; our best reason to deny P2, in other words, is our phenomenology of decision making. But if Libet's findings are correct, this phenomenology is *just plain wrong* at times; we can *feel* as if our conscious decision is causally-efficacious even when that decision has already been made on the neural level.[9] So then, while Libet's findings may

8. See, especially, Trevena & Miller (2010) amd Travena & Miller (2002).

9. See, though, Horgan (2011), who discusses the phenomenology of decision making in relation to Libet's findings.

not work as evidence for P2, they do undermine our best evidence for not-P2. Furthermore – and perhaps even more worrisome – Libet's findings suggests that empirical work in neuroscience eventually *could* confirm P2. So, rejecting P2 outright is a weak strategy in that it is open to empirical falsification.

In what follows, then, I target P1 – PCC. More specifically, I will argue that one can be held responsible for actions over which one does not have conscious control and that, therefore, PCC is false.

2.2.1 Counterexamples to PCC

According to PCC, we can be held responsible only for actions that are caused by a conscious decision. Any action which is not caused by a conscious decision but for which we still hold the agent morally responsible will thus function as an effective counterexample to the premise. But it turns out that finding such counterexamples is not very difficult. Consider, for instance, the following:

> **Gangster Wally**: Wally is a gangster, and has lived the gangster life so long that he acts like a gangster *without even thinking about it*. So, when Jerry – who runs the local roulette table – isn't looking, Wally switches out the normal dice with weighted dice out of habit, without consciously noticing what he is doing.
>
> **Saintly Wally**: Wally is walking through a poor section of town when a woman approaches him and asks for a meal. *Without thinking* and purely out of habit, Wally hands her the bag of Taco Bell he has just purchased.

Now, I take it that we want to hold Wally morally responsible for the actions described in each of these scenarios – it seems fully appropriate to blame Gangster Wally and to praise Saintly Wally.[10] However: *ex hypothesi* neither Gangster Wally nor Saintly Wally caused their respective actions by a conscious decision. The lack of conscious control in these scenarios, then, does not seem to affect the fact that we want to direct praise and blame at Wally. But if this is the case, the scenarios serve as counterexamples to PCC –

10. In fact, my own intuitions suggest that Wally is *more* praiseworthy and blameworthy in these scenarios than he would have been had he performed the actions by consciously deciding to do so. For isn't there something *especially* repugnant about the fact that Gangster Wally acts as he does without even thinking, and something *especially* praiseworthy about the fact that Saintly Wally acts as he does without even thinking? I won't belabor the point, as it is tangential to my immediate purposes, but it is, I think, worth thinking about.

they provide cases in which we hold agents morally responsible for actions even when these actions are not caused by a conscious decision.

2.2.2 Responses and Replies

Proponents of PCC might respond to this argument in a number of ways. First, they might argue that Wally cannot accurately be described as *acting* in these scenarios. Remember Velleman's position: since events that are not under one's conscious control are not *one's own*, these events cannot be considered actions. And certainly, if the events described in the scenarios above are not Wally's actions, he cannot be held morally responsible for them.

Now, a thorough reply to Velleman's position would involve a discussion of what does and does not count as an action. And the current paper is hardly the forum for *this*. Still, I want to suggest that while Velleman may accurately identify some strong sense of action, there is also a weaker sense of action, and that this weaker sense is also morally-relevant. For consider:[11] *first*, I take it that common sense counts behaviors caused by unconscious decisions as morally-relevant actions.[12] I take it, for instance, that the following would count as morally-relevant actions on an everyday understanding: my instinctually shoving an elderly lady in a rush to the subway; my habitually greeting my coworkers when I arrive in the morning; my deciding on a whim to ignore a colleague in the hallway. Insofar then as common sense treats such behaviors as morally-relevant actions, the burden lies with Velleman to show where common sense goes wrong. *Second*, it does not follow from a decision's being unconscious that the decision cannot be understood in terms of reasons. Saintly Wally is a good example of this. For it isn't that his decision to give away his Taco Bell is irrational or mysterious; Saintly Wally, I take it, clearly does have reasons for his behavior and could articulate these reasons if he wanted to. It's just that these reasons remain unconscious and unarticulated. But if Saintly Wally's behaves as he does *for a reason*, this counts in favor of his behavior counting as a morally-relevant action on many theories.[13] *Third*, many significant moral decisions are made unconsciously. As Arpaly (2003) suggests, many of us have "made a drastic career change, left a marriage, a

11. These arguments that follow largely follow those put forward by Arpaly (2003).

12. Arpaly (2003) points out – correctly, I think – that moral philosophy can skew our understanding of what counts as an action. The typical agent of moral philosophy, after all, is a conflicted deliberator. But this obscures the fact that *most* of our actions are carried out habitually, not consciously or deliberatively.

13. Scanlon (2008), among others, argues for the relevance of reasons to action. See, for instance, pages 122–131.

church, or a cult, or otherwise made a hard choice" without consciously deciding to do so; we have, rather, *found ourselves* making or having made these decisions (Arpaly 2003, 4). But certainly, if anything should count as an action for which we can be held responsible, important moral decisions such as these should. And if these decisions count as morally-relevant actions, then morally-relevant actions can be caused unconsciously.

Suppose that the proponent of PCC finds these arguments convincing. There is a still another strategy available to her. In particular, she can reason as follows: sure, we want to hold Wally morally responsible in the scenarios presented above. But it isn't that we hold him morally responsible for *what he does*. Rather, we hold him morally responsible for *who he is*. When we praise Saintly Wally and blame Gangster Wally, in other words, we are evaluating Wally as, respectively, a good guy and a bad egg. And this blame or praise is in turn justifiable just in case the actions that led Wally to become who he is were the result of a conscious decision.[14] When we hold Wally responsible in the scenarios presented above, we are thus *ultimately* holding Wally responsible for actions that were the result of a conscious decision; it's just that these actions happened in the past. But if this is the case, the scenarios do not work as effective counterexamples to PCC.

In response: if we suppose that Wally is morally responsible *solely* for past actions, we thereby commit ourselves to the position that Wally's present actions are irrelevant to our praise and blame. But this is problematic. For suppose that Gangster Wally hadn't switched out the dice. If our blame of Gangster Wally is directed exclusively at his past actions, then the Gangster Wally who *doesn't* switch out the dice would be just as blameworthy as the Gangster Wally who does – the two Gangster Wallys, after all, would have performed identical past actions. But this seems obviously wrong. For it seems obvious that the Gangster Wally who switches out the dice is more blameworthy than the one who does not. And the explanation for this also seems obvious: when we blame Gangster Wally, part of what we are doing is blaming him for his *present* action. It is therefore problematic to claim that we praise and blame Wally solely for those actions that led Wally to become who he is.

Proponents of PCC, however, could advance still a third objection. In particular, they could emphasize just how *intuitive* PCC is. PCC, after all, seems able to account for why we don't hold people accountable for twitches, schizophrenic episodes, compelled misdeeds and so on – it seems, after all, that we do not hold people responsible for such actions precisely because these actions are not caused by a conscious decision. At the very least,

14. So, if Wally *didn't* consciously choose to become the way he is, we *wouldn't* be justified in holding him morally responsible. Wolf (1990)'s figure Jojo offers an occasion for thinking about this.

this objector could continue, one should be able to account for why we often make what seem to be fully appropriate counterfactual claims of the form: "had S caused *a* by way of conscious decision, then S would have been morally responsible *a*." But it isn't clear, or so the objection goes, how one can hold on to such claims if one rejects PCC. By rejecting PCC, in short, one loses sight of the obvious: conscious control is importantly relevant to moral responsibility.

I sympathize with this worry, as I do suspect that conscious control is importantly relevant to our moral responsibility practices. But I don't think the objection is damaging to my position. This is for two reasons: *first*, all I have argued here is that conscious control is not a *necessary* condition for holding an agent morally responsible. There is thus nothing I have said that would make it problematic to contruct an account according which conscious control (along with some other conditions, no doubt) is *sufficient* for holding an agent morally responsible. And such an account of moral responsibility, it seems, could more than adequately deal with the intuitive pull of counterfactuals of the form above. *Second*, nothing I have said entails that conscious control does not importantly affect our moral responsibility practices. It is perfectly compatible with everything I have said, for instance, that Gangster Wally would have been more (or less) morally responsible for switching out the dice had he done so consciously.

In short: on a complete account of our moral responsibility practices, it may well be that conscious control is not only relevant, but central. But there is nothing I have said that is incompatible with such an account; all I have tried to show here is that there are actions which are not caused by a conscious decision but for which an agent can still be held morally responsible. For if this is the case, PCC is false, and the argument from Libet's findings presented in section 2.1 does not go through.

3.0 Neuroscience as Evidence for Exemptions

Given what I have argued above, one might wonder: are there *any* ways in which recent advances in neuroscience are importantly relevant to our moral responsibility practices? In the current section, I want to suggest that there are, but that neuroscience finds its proper place in these practices only after some serious philosophical work has been completed. More specifically, I will suggest that neuroscience can be used as evidence for deciding when individuals meet philosophically-defined criteria for moral exemption.

3.1 Neuroscience as Non-Sufficient Evidence for Moral Exemption

Jay Wallace (1994) argues that we should, in certain circumstances, exempt particular individuals from our more responsibility practices, especially our practices of praise and blame. More specifically, Wallace argues that we should exempt any agent from moral responsibility who lacks "the power to grasp moral reasons and the power to control their behavior in accordance with them" (Wallace 1994, 162). Wallace then gives some examples of agents who seem to meet this criterion: young children; individuals afflicted with mental illness; agents under hypnosis; and so on. According to Wallace, these individuals either cannot fully grasp moral reasons or cannot control their behavior in accordance with them, and should not, therefore, be subject to normal moral practices of praise and blame (Wallace 1994).

Now, there are other accounts of moral exemption, and taking sides in *this* debate is not among my purposes here. Rather, I want to point out two features of Wallace's methodology that I find helpful. *First*, Wallace seems right that it is important for us to exempt at least some individuals from our normal practices of praise and blame – clearly, we do not want to blame and praise children and those with schizophrenia in the same way we praise and blame normally-functioning adults. When we praise and blame normally functioning adults, after all, we think of this as a form of deserved condemnation; when we praise and blame children or those with schizophrenia, on the other hand, we think of this as a form of education, protection, adjustment, or regulation. *Second* – and as Wallace also sees – if we want to exempt certain individuals from moral responsibility, it becomes an important philosophical project to determine what the criteria for exemption are. It becomes an important project, in other words, to determine some principled way of deciding when an agent should be held exempt from moral responsibility. This is no small project. And it is, it must be emphasized, a distinctively *philosophical* one. For it seems obvious that any account of moral exemption will have to reference standards of rationality, the appropriateness of our moral responsibility practices, the nature of moral reasons, our ability to grasp these reasons, or some such criteria. And it isn't immediately clear that empirical findings have much to say about any of this.

Once we determine the criteria for moral exemption, however, empirical input becomes important. For suppose that, i.e., we adopt Wallace's criteria for moral exemption – suppose that an inability to grasp moral reasons is sufficient for exemption from moral responsibility. And further suppose that anyone diagnosed with schizophrenia is unable to grasp moral reasons. How then do we determine which individuals are schizophrenic? Clearly, by way of empirical evidence: behavioral analyses, case histories, expert testimony, and so on. And here too is where contemporary neuroscience finds a place of relevance in regard to our moral responsibility practices. Neuroscience has, after all, successfully

located specific neural correlates of schizophrenia, as well as neural correlates of other potentially exempting disorders (Hyde & Weinberger 1990; Torrey et al. 1994; Suzuki et al. 2002; Thompson et al. 2001; Bremner et al. 1995; Gilbertson et al. 2002). So, if we can determine that an individual has one of these neural correlates, this gives us evidence for thinking that the individual should be diagnosed with the relevant disorder. And this, in turn – again, assuming that a diagnosis of the relevant disorder is sufficient evidence that philosophically-determined criteria for moral exemption have been met – could give us reason to exempt the individual from our normal moral responsibility practices.

A word of caution, however: both psychologists and neuroscientists consistently warn how easy it is to become enamored with neuroscientific evidence.[15] For it is all too easy to forget that psychological disorders are defined *behaviorally*; they are not defined in neuronal terms – on its own, then, neuroscientific evidence can never offer conclusive evidence for the diagnosis of any disorder. If it's not clear why this is the case, think about it this way: an individual who exhibits cortical activation in the area of primary motor cortex that is mapped to hand movement but who is *not moving her hand* simply cannot be described as moving her hand. Similarly, an individual who exhibits the neural correlates of schizophrenia but who is behaviorally normal simply cannot be described as schizophrenic. This thus places an important limitation on the relevance of neuroscientific evidence to our moral responsibility practices. In particular, even if being diagnosed with a certain disorder is sufficient for being held exempt from our moral responsibility practices, and even if there is neural evidence that can help make this diagnosis, this evidence can never be taken as sufficient for a diagnosis. Therefore, the neural evidence relevant to moral exemption should never be treated as sufficient evidence for moral exemption.

So, let's take stock. Neuroscience, it turns out, *is* relevant to our moral responsibility practices. In particular, neuroscientific evidence is relevant to our moral responsibility practices as long as the following conditions are met:

> **Criterion Relevancy** – the evidence is framed as providing reason that some philosophically-defined criterion for moral exemption has been met.
>
> **Non-Sufficiency** – the evidence is treated as *non-sufficient* evidence for moral exemption.

15. See, for instance, Morse (2006) and Racine et al. (2005).

These are strict conditions. They may, in fact, lead one to conclude that neuroscience isn't all that relevant at all to our moral responsibility practices, after all. And in a way, this is the thesis of the current paper. Contrary to what some have argued and to the way it may seem, neuroscience *isn't* very relevant to our moral responsibility practices; it finds its proper role only after a significant amount of philosophical work has been completed. And even then, its role is rather humble.

4. Conclusion

In this paper, I have tried to show how neuroscience is and is not relevant to our moral responsibility practices. More specifically: I have argued that – contrary to what some have supposed – the mechanistic explanations of human action offered by neuroscience are not capable on their own of leading us to revise our moral responsibility practices (Section 1). Next, I argued that while findings such as those of Libet et al. (1983) may *seem* to universally exempt us from moral responsibility, universal exemption does not in fact follow from these findings (Section 2). I concluded, however, that neuroscience is relevant to our moral responsibility practices as long as it is properly situated, as long as it is used as non-sufficient evidence that philosophically-defined criteria for moral exemption have been met (Section 3).

References

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Ayer, A. J. 1954. *Philosophical Essays*. New York: Macmillan.

Banks, William P. and Eve A. Isham. 2011. "Do We Really Know What We are Doing? Implications of Reported Time of Decision for Theories of Volition." In *Conscious Will and Responsibility*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 47–61. Oxford: Oxford University Press.

Bayne, Tim. 2011. "Free Will and the Sciences of Human Agency." In *Free Will and Modern Science*, edited by Richard Swinburne. Oxford: Oxford University Press.

Bechtel, William and Robert E. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Breedlove, S.M., N.V. Watson, and M.R. Rosenzweig. 2010. *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience* (6th ed.). Sunderland, MA: Sinauer Associates, Inc. Publishers.

Bremner, J.D., P. Randall, T.M. Scott, R.A. Bronen, et al. 1995. "MRI-Based Measurement of Hippocampal Volume in Patients with Combat-Related Posttraumatic Stress Disorder." *American Journal of Psychiatry* 152 (7): 973–981.

Chisholm, Roderick. 1964. "Human Freedom and the Self." In *Free Will,* edited by Robert Kane, 47–58. Malden: Blackwell.

Fischer, John Martin. 1987. "Responsiveness and Moral Responsibility," In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Shoeman, 81–106. Cambridge: Cambridge University Press.

Fischer, John Martin. 1994. *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.

Ekstrom, Laura W. 2000. *Free Will: A Philosophical Study*. Boulder: Westview.

Farah, Martha J. and Andrea S. Heberlein. 2007. "Personhood and Neuroscience: Naturalizing or Nihilating?" *The American Journal of Bioethics* 7 (1): 37–48.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20.

Gilbertson, M.W., M.E. Shenton, A. Ciszewski, K. Kasai, et al. 2002. "Smaller Hippocampal Volume Predicts Pathologic Vulnerability to Psychological Trauma." *Nature Neuroscience* 5 (11): 1242-1247.

Greene, J. & J. Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences* 359 (1451): 1775–1785.

Hallett, Mark. 2011. "Volition: How Physiology Speaks to the Issue of Responsibility." In *Conscious Will and Responsibility*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 61–69. Oxford: Oxford University Press.

Haggard, P. and M. Eimer. 1999. "On the Relation Between Brain Potentials and the Awareness of Voluntary Movements." *Experimental Brain Research* 126 (1): 128–133.

Haynes, John-Dylan. 2011. "Beyond Libet: Long-Term Prediction of Free Choices from Neuroimaging Signals." In *Conscious Will and Responsibility*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 85–96. Oxford: Oxford University Press.

Horgan, Terry. 2011. "The Phenomenology of Agency and the Libet Results." In *Conscious Will and Responsibility*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 159–172. Oxford: Oxford University Press.

Hyde, T. M. and D. R. Weinberger. 1990. "The Brain in Schizophrenia." *Seminars in Neurology* 10 (3): 276–286.

Kane, Robert H. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Determinism." *Journal of Philosophy* 96 (5): 217–40.

Keller, I. and H. Heckhausen. 1990. "Readiness Potentials Preceding Spontaneous Motor Acts: Voluntary vs. Involuntary Control." *Electroencephalography and Clinical Neurophysiology* 76 (4): 351–361.

Lau, H. C., R.D. Rogers, P. Haggard, and R.E. Passingham. 2004. "Attention to Intention." *Science* 303 (20): 1208–1210.

Libet, Benjamin. 2011. *Do We Have Free Will?* In *Conscious Will and Responsibility*, edited by Walter Sinnott-Armstrong and Lynn Nadel, 1–10. Oxford: Oxford University Press.

Libet, B., C. Gleason, E. Wright, and D. Pearl. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initiation of a Freely Voluntary Act." *Brain* 106 (3): 623-664.

Morse, Stephen J. 2006. "Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note." *Ohio State Journal of Criminal Law* 2 (397): 397–412.

Nahmias, Eddy A., Stephen G. Morris, Thomas Nadelhoffer, and Jason Turner. 2005. "Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility." *Philosophical Psychology* 18 (5): 561–584.

Nahmias, Eddy, D. Justin Coates, and Trevor Kvaran. 2007. "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions." *Midwest Studies in Philosophy* 31 (1): 214–242.

Nichols, Shaun and Joshua Knobe. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Noûs* 41 (4): 663–685.

Pereboom, Derk. 2013. "Free Will Skepticism, Blame, and Obligation." In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini, 189–206. Oxford: Oxford University Press.

Pockett, Susan. 2004. "Does Consciousness Cause Behaviour?" *Journal of Consciousness Studies* 11 (2): 23–40.

Racine, E., O. Bar-Ilan, and J. Illes. 2005. "fMRI in the Public Eye." *Nature Reviews Neuroscience* 6 (2): 159–164.

Roediger, H. L., M. K.Goode, and F.M. Zaromb. 2008. "Free Will and the Control of Action." In *Are we Free? Psychology and Free Will*, edited by J. Baer, J. C. Kaufman, and R. F. Baumeister, 205–225. Oxford: Oxford University Press.

Scanlon, Thomas. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge: Belknap Press.

Soon, C.S., M. Brass, H.J. Heinze, and J.D. Haynes. 2008. "Unconscious Determinants of Free Decisions in the Human Brain." *Natural Neuroscience* 11 (5): 543–545.

Spence, Sean. 2009. *The Actor's Brain: Exploring the Cognitive Neuroscience of Free Will*. Oxford: Oxford University Press.

Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1-25.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

Suzuki, M., S. Nohara, H. Hagino, K. Kurokawa, et al. 2002. "Regional Changes in Brain Gray and White Matter in Patients with Schizophrenia Demonstrated with Voxel-Based Analysis of MRI." *Schizophrenia Research* 55 (1–2): 41–54.

Thompson, P.M., J.N. Giedd, R.P. Woods, D. MacDonal, et al. 2000. "Growth Patterns in the Developing Brain Detected by Using Continuum Mechanical Tensor Maps." *Nature* 404 (6774): 190–193.

Torrey, E.F., A.E. Bowler, E.H. Taylor, and I.I. Gottesman. 1994. *Schizophrenia and Manic Depressive Disorder*. New York: Basic Books.

Trevena, J.A. and J. Miller. 2002. Cortical Movement Preparation Before and After a Conscious Decision to Move. *Consciousness and Cognition* 11 (2): 162–190.

Trevena, Judy and Jeff Miller. 2010. "Brain Preparation Before a Voluntary Action: Evidence Against Unconscious Movement Initiation." *Consciousness and Cognition* 19 (1): 447–456.

van Inwagen, Peter. 1975. "The Incompatibility of Free Will and Determinism." *Philosophical Studies* 27: 185–199.

Velleman, D.J. 1992. "What Happens When Someone Acts?" *Mind* 101 (403): 461–481.

Wallace, R. Jay. 1996. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–248.

Wegner, Daniel M. 2003. *The Illusion of Conscious Will*. Cambridge: MIT Press.

Wiggins, David. 1973. "Towards a Reasonable Libertarianism." In *Essays on Freedom of Action*, edited by Ted Honderich, 31. New York: Routledge.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Wright, R. 1994. *The Moral Animal: Evolutionary Psychology and Everyday Life*. New York: Pantheon.

Journal of Cognition and Neuroethics

Review of *Neuroethics in Practice: Medicine, Mind, and Society*

James Beauregard
Rivier University

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). November, 2014. Volume 2, Issue 2.

Citation
Beauregard, James. 2014. "Review of *Neuroethics in Practice: Medicine, Mind, and Society*." *Journal of Cognition and Neuroethics* 2 (2): 83–87.

Review of *Neuroethics in Practice: Medicine, Mind, and Society*

James Beauregard

Chatterjee, Anjan, and Martha J Farah, eds. 2013. *Neuroethics in Practice: Medicine, Mind, and Society*. New York: Oxford University Press. 278 pages. $63 hardcover. $38.99 Kindle edition.

Neuroethics in Practice: Medicine, Mind, and Society brings together authors from several disciplines writing on contemporary issues at the intersection of neuroscience and bioethics. The book's editors, Anjan Chatterjee, Professor of Neurology and a member of the Center for Cognitive Neuroscience and the Center for Neuroscience at Society at the University of Pennsylvania, and Martha Farah, Walter H Annenberg professor of Natural Sciences at the University of Pennsylvania and director for the Center for Neuroscience and Society, (and the editor of the 2010 *Neuroethics: An Introduction with Readings*), know the territory and have published extensively in the field. The book highlights many of the strengths of the burgeoning field of Neuroethics as well as some noteworthy weaknesses.

Neuroethics in Practice is divided into five sections, each focusing on a major issue in contemporary neuroethical thinking. In the first section, "Brain Enhancement," Chatterjee addresses the issue of "cosmetic neurology," the practice of attempting to enhance aspects of functioning and healthy individuals, including anabolic steroids and autologous blood transfusions in professional athletes, the medications traditionally used for treatment of Attention Deficit Disorder, and a more recent use of cholinesterase inhibitors to attempt to improve memory in the healthy; and for improvement in mood, St. John's Wort and selective serotonin reuptake inhibitors. Chatterjee organizes his presentation around four ethical issues; safety, distributive justice, coercion and potential erosion of character, noting that the notion of weighing risks and benefits traditional to medicine cannot be applied in the same manner to healthy individuals seeking enhancement. He touches on notions of character and personhood and the "fundamental concern that chemically changing the brain threatens our notion of personhood" through diminishing the adversity that builds character, or, in his example, considering the manner in which dampening painful memories "change who we are, if we are to some degree the sum of our experiences" (8). Enhancement also raises obvious questions of distributive justice, in terms of who has

access to such enhancers, and who will pay for them. Lastly, he raises the ethical issue of potential coercion to use pharmacological enhancement both in higher education and the workplace.

Ilina Singh and Kelly Kelleher addressed the thornier issue of neuroenhancement in the young, asking if cognitive enhancement is warranted "where cognitive and/or behavioral functioning is not judged to be impaired" (17), discussing how our notion of impairment has become more fluid with the advent of potential for enhancement. There also raise the concern that there is little long-term data available about enhancement use when neurodevelopment is still in progress, and the related, clinically and legally complex question of the extent to which those under the age of 18 can consent to or dissent from neuroenhancement use. Accepting that neuroenhancement is here to stay, they recommend that the primary care physician be the gatekeeper for such interventions.

Cognitive enhancement in the military is addressed by Michael B Russo, Melba C. Stetz and Thomas A. Stetz, who note that caffeine and nicotine have long been used as enhancements in the context of military service. They raise issues of potential coercion and raise the potential of what they term "command–dictated preventative medicine" in high risk military situations, asking whether a determination might be made by the command structure that individual rights of soldiers might be superseded by the good of the whole if cogniceuticals are deemed likely to enhance group survival in high risk situations.

Direct to Consumer Advertising (DTC) is addressed by Peter Conrad and Alan Horwitz, who note that concepts of illness, disease, and health have begun shift as pharmaceuticals have been marketed to the general public. In this connection, Breehan Chancellor and Anjan Chatterjee review the research on brain training programs (such as Lumosity) and the manner in which neuroscientific data can be manipulated in marketing strategies, noting that "commerce has moved ahead of its science" (61).

The book's second section looks at issues of Competence and Responsibility, beginning with how both are understood in current clinical and legal contexts as they touch on issues of driving, voting and financial decision-making (Jason Karlawish), informed consent for treatment and research (Scott Y.H. Kim), and legal issues of responsibility in cases of addiction (Steven E. Hyman), noting the increasing importance of neurocognitive data in these decision-making processes.

Brain Imaging is reviewed in the book's third section, beginning with medical/legal issues in the use of neuroimaging data (Stacey Tovino), who notes that neuroimaging, which might initially be seen as a straightforward clinical process, has entered into criminal, civil and administrative law. She traces lawsuits that are filed for negligent

neuroimaging (filed by patients or families who are been injured or killed in the course of MRI scanning that falls below standards of practice), and the use of neuroimaging data in assessing disorders of consciousness. John Detyre and Tamara Bockow address the ethical issues around what should be done with incidental findings in MRI research, noting that highly specific research studies often do not meet the clinical standards for neuroimaging and that clinical significance of neuroimaging data is not always clear, two important factors in determining when or how finding should be disclosed to subjects in research studies. Martha Farah and Seth J. Gillihan write a fascinating chapter on the uses of neuroimaging in clinical psychiatry, noting its potential to be used diagnostically though such utility at this point in time has a limited scientific foundation. They present the case of the Amen Clinics founded by psychiatrist Daniel Amen (http://www.amenclinics.com), which claims to offer SPECT scanning for reliable psychiatric diagnosis, reviewing the research on neuroimaging, and concluding that this is a problematic and at best unproven technology for the psychiatric realm in need of further study before real diagnostic utility can be achieved. Not mentioned in this book, but equally concerning, is the use of neuroimaging data in forensic settings, typified by *No Lie MRI* (http://www.noliemri.com), which claims better lie detection via neuroimaging technology for use in forensic settings.

The books fourth section touches on the issue familiar to readers who delve into bioethics, severe brain damage, examining current neurologic criteria for brain death (Steven Laureys), vegetative state and minimally conscious state (Joseph Fins and Nicholas D. Schiff), who also examine the ways in which neuroimaging data has exerted a role in improving diagnostic clarity, and wonder if this technology might one day "provide a mechanism for communication" for injured individuals unable to communicate through normal channels.

Martha Farah provides the book's only sustained reflection on our notions of personhood, and notes means by which neuroscience data can demonstrate "preserved capacity for cognition, consciousness and interpersonal communication" in brain-damaged individuals (185).

The final section, titled "New Treatments, New Challenges" is excellent overview of cutting-edge technology in the treatment of neurologic illness including Deep Brain Stimulation (Matthis Synofzik), Transcranial Magnetic Stimulation (Horath, Perez, Forrow, Frengi and Pascual-Leone), Implanted Neural Interfaces, "devices which interface directly with the nervous system for the monitoring and modulation of neural tissue," including deep brain stimulators, spinal cord stimulator's and epilepsy monitoring and

suppression systems; they also address issues of informed consent in neurologically impaired individuals.

The book concludes with a brief chapter titled "Biologics in the Human Brain" in which Jonathan Kimmelman views the sciences of gene transfer, cell transplantation and neurotrophic factors for treating neurologic injury and disease both in research and clinical practice.

Neuroethics In Practice and shares the strengths and weaknesses of many contemporary works in the field. Such works are typically long on the neurologic and technical aspects but short on the philosophical/bioethical dimension of contemporary neuroscience. This shortfall often manifests itself in two ways. First, Neuroethics is not gone very far in developing comprehensive, nuanced visions of person and personhood. The attempts that are made typically move from a reductionist scientific perspective, equating person with function, typically expressed in the context of severe brain damage, with its profound practical implications around determination of death and organ transplantation. Second, despite the vast technological advances that have occurred in neuroscience, the field lacks any developed philosophy of technology despite extensive reflection in this area by thinkers as varied as Martin Heidegger, Herbert Simon, Michael Polanyi, Peter-Paul Verbeek, Hans Jonas and many others.

With these limitations in mind, *Neuroethics In Practice* is an excellent introduction to the clinical and technological advances made possible by neuroscience, providing a solid foundation for those who wish to enter into conversation about the ethical questions these new technologies raise.

cognethic.org