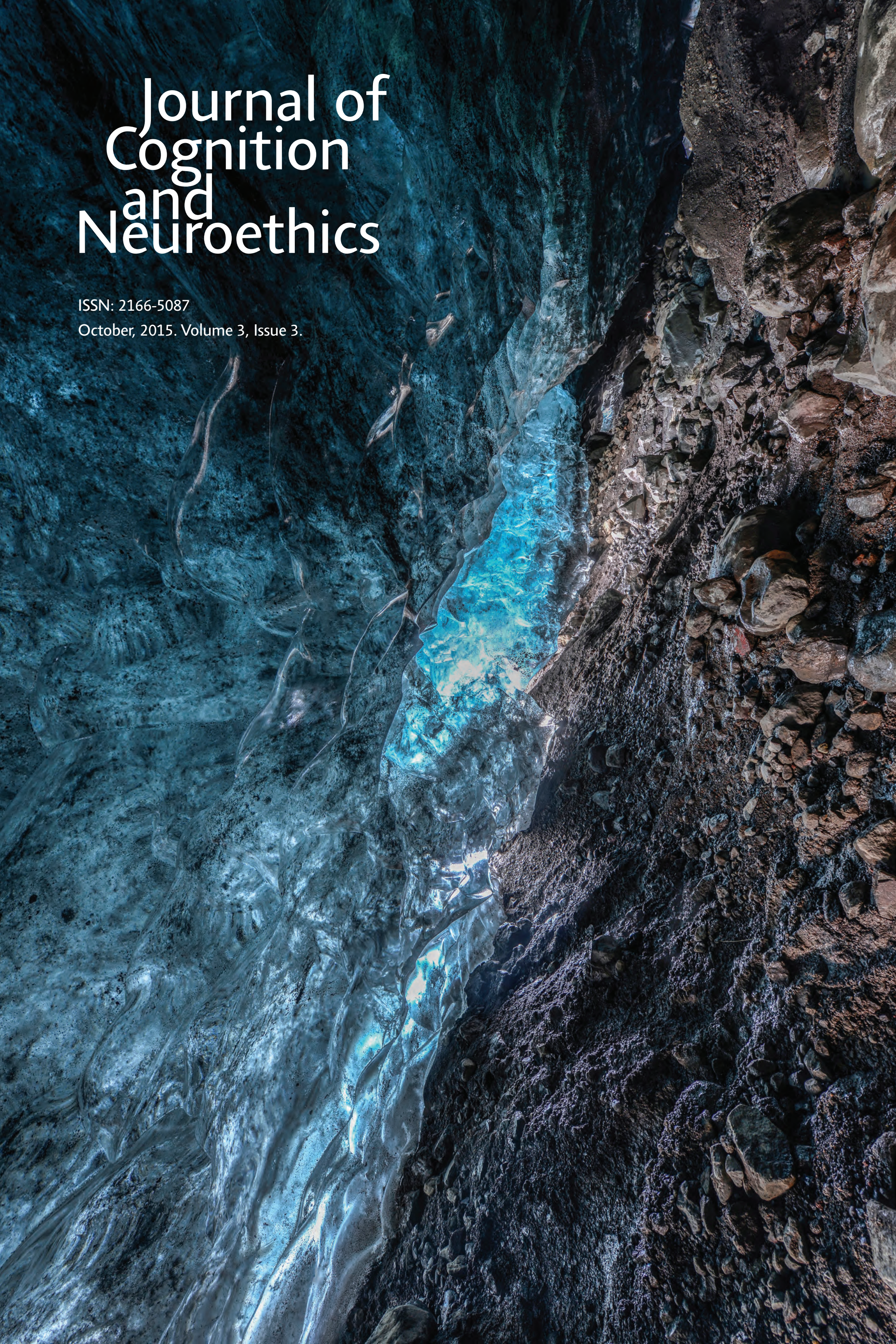

Journal of Cognition and Neuroethics

ISSN: 2166-5087
October, 2015. Volume 3, Issue 3.

Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Volume 3, Issue 3 Special Issue Editor
Zea Miller

Production Editor
Zea Miller

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

© 2015 Center for Cognition and Neuroethics

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

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

Table of Contents

Introduction

Science fiction is obsessed with brains. Through modeling various forms of cognition from the advanced (e.g., *Xeelee Sequence*) to the extraordinary (e.g. *Solaris*) and from the artificial (e.g., *2001: A Space Odyssey*) to the disembodied (e.g., *The City and the Stars*), the genre explores, though sometimes conservatively, what is ideally human. By engaging our minds with moral questions ranging from the use of telepaths as weapons (e.g. *Babylon 5*) to developing neurological WMDs (e.g., *Star Trek: Voyager*) and from perfection (e.g. *GATTACA*) to enhancement (e.g., *Lucy*), the genre examines what William Safire defines as neuroethics: "what is right and wrong, good and bad about the treatment of, perfection of, or unwelcome invasion of and worrisome manipulation of the human brain."[1] The work of cognition and neuroethics in science fiction therefore advances or articulates cultural theories and ideals, from neurochauvinism and negative eugenics[2] to saving the world and the meaning of life. Unsurprisingly, the tension between the two are as startling and complex as the societies from which they emerged.

While research on the relationship between cognition and reading science fiction is extensive,[3] the ways in which cognition and neuroethics are narratologically deployed in these narratives, and to what end, remains relatively unexamined – so much so that I felt we had a groundbreaking opportunity. So, in 2014, the Center for Cognition and Neuroethics (CCN) – a joint venture between the Philosophy Department at the University of Michigan-Flint and the Insight Institute of Neurosurgery and Neuroscience (IINN) – issued a call for proposals for its first conference on cognition and neuroethics of science fiction. The goal was to foster a space where scholars from multiple disciplines could discuss the work of cognition and neuroethics in science fiction, and thereby begin a new critical conversation – and it worked. In March of 2015, CCN brought 16 scholars together at IINN for 6 panel discussions on the theme. All participants were invited to submit their revised talks for consideration in this special issue. After peer reviews and revisions, we invited to include the 6 articles you find published herewith.

In this special issue, contributors examine the work of neuroethics and cognition in morality, subjectivity, consent, and thought experiments. By leveraging cognition from

[1] Safire, William. 2002. "Visions for a New Field of 'Neuroethics.'" *Neuroethics: Mapping the Field Conference Proceedings*. San Francisco, California: May 13–14.

[2] After Anderson, Jami. 2012. "A Dash of Autism." In *The Philosophy of Autism*, edited by Jami Anderson and Simon Cushing, 109–142. Lanham: Rowman and Littlefield.

[3] See Suvin, Darko. 1979. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. New Haven, CT: Yale University Press.

philosophy of mind and neuroethics from bioethics, these articles represent a neuroethical turn for science fiction studies and philosophy of literature, as a new transdisciplinary third space. Together, they create a grounding for the neuroethics of science fiction as a field of inquiry. It is my hope that, by interrogating the functions and operations of neuroethical work in science fiction, the cultural production of ideals and humanity that are entangled with cognition in the genre will, in time, be revealed with greater, nuanced granularity.

Zea Miller
Special Issue Editor
Journal of Cognition and Neuroethics

Journal of Cognition and Neuroethics

Apes with a Moral Code? Primatology, Moral Sentimentalism, and the Evolution of Morality in *The Planet of the Apes*

Paul Carron
Baylor University

Biography

Paul Carron is an Assistant Professor of Philosophy in the Baylor Interdisciplinary Core at Baylor University, where he teaches courses at the intersection of philosophy, psychology, ethics, and social theory. His primary research interests are ethics, free will/moral responsibility, moral psychology, and ancient philosophy.

Acknowledgements

I would like to thank Leslie Roy Ballard for his very helpful comments and editing suggestions on the earlier drafts of this paper. I would also like to thank the members of my writing group, Anne-Marie Schultz, Tom Hanks, and Davide Zori for helpful comments and suggestions. Finally, thanks to Zea Miller for organizing a fine conference and to the participants at the Work of Cognition and Neuroethics in Science Fiction conference at the Insight Institute of Neurosurgery and Neuroscience.

Publication Details

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

Citation

Carron, Paul. 2015. "Apes with a Moral Code? Primatology, Moral Sentimentalism, and the Evolution of Morality in *The Planet of the Apes*." *Journal of Cognition and Neuroethics* 3 (3): 1–26.

Apes with a Moral Code? Primatology, Moral Sentimentalism, and the Evolution of Morality in *The Planet of the Apes*

Paul Carron

Abstract

This essay examines the recent *Planet of the Apes* films through the lens of recent research in primatology. The films lend imaginary support to primatologist Frans de Waal's evolutionary moral sentimentalism; however, the movies also show that truly moral emotions outstrip the cognitive capacities of the great apes. The abstract moral principles employed by the ape community in the movie require the ability to understand and apply a common underlying explanation to perceptually disparate situations; in contrast, recent research in comparative psychology demonstrates that the great apes lack this capacity. Since the capacity for abstraction is required on even the most basic version of moral sentimentalism—Shaun Nichols' sentimental rules account—the lack of the capacity for abstraction reveals a qualitative distinction between primate social behavior and human morality.

Keywords

Moral Sentimentalism, Impartial Spectator, Frans de Waal, Primatology, *Planet of the Apes*

> Any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well developed, or nearly as well developed, as in man. (Darwin [1871] 1981, 71–72)

> To name an act good or bad, ultimately implies that it is apt to give rise to an emotion of approval or disapproval in him who pronounces the judgment… (Westermarck 1906, 4)

> Ape separate weak. Ape together strong… Ape not kill ape. (Wyatt 2011; Reeves 2014)

Introduction

The Planet of the Apes is a classic American science fiction film. Released in 1968, it received critical acclaim and box office success ("Top-US-Grossing" 2015). Its inclusion in the National Film Registry by the Library of Congress for cultural, historical, and aesthetic significance evidences its lasting impact. The original film spawned four sequels and several other spinoffs. This film captivated audiences not only for its technical achievements, such as its cinematography, realistic ape costumes, and haunting score, but also for the interesting philosophical questions it raised.

In 2001, a reboot of the franchise appeared in theaters—*The Planet of the Apes* directed by Tim Burton—that had relative box office success but failed critically and did little to advance the franchise's more interesting philosophical and cultural importance. Another reboot of the franchise appeared in 2011. Thus far, the most recent reboot has generated two films, *The Rise of the Planet of the Apes* and *Dawn of the Planet of the Apes*. Whereas the original film tries to debunk various aspects of human exceptionalism by exposing the human propensities toward racism, classism, and religious ideology, *Rise* and *Dawn* imaginatively depict great apes developing human capacities such as reflective consciousness, emotional intelligence, and moral reasoning.[1]

Recent advances in primatology and comparative psychology reveal that many of those imaginative depictions are not so fanciful; rather, the portrayals reflect our growing understanding of nonhuman primate cognitive capacities, an understanding due in large part to the careful research, spanning over 30 years, of primatologist Frans de Waal. Putting the recent reboots into dialogue with current scientific research like de Waal's illuminates the extent to which human morality is rooted in primate social behavior and cognitive capacities. However, careful examination of that research and its philosophical framework reveals how human moral reasoning and action far outstrip the capacities of nonhuman primates.

In particular, examining what I label de Waal's evolutionary moral sentimentalism sheds light on the fictitious story of the reboots. Briefly, moral sentimentalism is the idea that moral judgments stem from a person's reactive attitudes—sympathy, anger, compassion, resentment, etc. When a person judges that an action is morally wrong, this judgment happens in part because the agent experiences—or believes she should experience—certain reactive attitudes in response to the action in question.

Despite evidence that supports de Waal's evolutionary moral sentimentalism, certain aspects of the reboots go beyond de Waal's theory to reveal a fundamental problem with

[1] Whenever I refer to the "reboots," I am referring to *Rise* and *Dawn*.

his account. The reboots portray the apes as developing the capacity for abstract moral reasoning, seen in their ability to adopt and apply abstract moral principles. De Waal does not explain how the common ancestor of the great apes and humans developed this capacity. Furthermore, recent research suggests that, notwithstanding all their profound intellectual capacities, the great apes lack the capacity for abstract moral reasoning. Even given all that we have in common with our primate cousins, a *qualitative* difference remains between primate social behavior and human morality. I conclude that despite its explanatory power, de Waal's theory fails to account for the evolution of a distinctly human morality.

My argument proceeds as follows. In section one, I briefly outline the tenets of a recent novel articulation of moral sentimentalism, as it is relevant to the films and for de Waal's primatology. In section two, I briefly overview the plot of the reboots, stressing scenes that display how the moral community evolves. The movies emphasize, for example, the importance of emotions and abstract moral principles for moral community, themes that later reemerge when discussing de Waal's work. In section three, I discuss what I define as de Waal's evolutionary moral sentimentalism (EMS). De Waal's EMS has two main features: 1) reactive attitudes, such as sympathy and empathy; and 2) moral norms that contribute to communal harmony. Major elements of de Waals' EMS are highlighted in the films.

Finally, section four contains the crux of my argument. De Waal argues that human morality evolved from primate social behavior, a theory creatively depicted in the reboots. However, the movies make clear a crucial aspect of human morality: abstract moral principles. De Waal also understands the importance of abstraction when he discusses the philosopher Adam Smith's notion of the "impartial spectator"—the view that when making moral decisions, we must abstract ourselves from our particular viewpoint to see the situation as an impartial person would see it. However, de Waal's account fails to provide evidence of how nonhuman primates develop a similar capacity for disinterestedness or applying abstract rules to the community as a whole. The ability to conceptualize and verbalize abstract notions differs radically between primates and humans. Even a more restrained "sentimental rules" account requires an agent to be able to distinguish between morals and conventions. Further, the agent must be able to recognize generalizable norms or principles regarding harm; recent research offers, however, that nonhuman primates lack the capacity for abstraction or generalization. I conclude that a qualitative distinction remains between primate social behavior and human morality.

Section One: The "Sentimental Rules" Account of Moral Sentimentalism

I am arguing that the most recent *Planet of the Apes* films present an imagined version of Frans de Waal's evolutionary moral sentimentalism. Therefore it is first necessary to understand the ethical theory referred to as moral sentimentalism. At the most basic level, moral sentimentalism is the ethical theory that ethical responses stem from emotions and feelings (de Waal 2006, 18). Sentimentalism stresses "empathetic caring as the touchstone of virtuous agency" (Cox 2006, 506). As the philosopher David Hume puts it, "The final sentence . . . which pronounces characters or actions amiable or odious, praiseworthy or blameable . . . depends on some internal sense or feeling which nature has made universal in the whole species" ([1777] 1975, 172–173).[2]

Contemporary neosentimentalists link moral judgment inextricably to *appropriate* moral feelings, primarily empathy, and the agent's ability to discern the appropriate moral feeling and act in accordance with that feeling in the absence of said feeling (D'Arms and Jacobson 2000, 729).[3] Moral sentimentalism has many varieties, but, for my purposes, I focus on one recent novel version, Shaun Nichols' "sentimental rules" account. I focus on this account because of its power to shed explanatory light on the films and de Waal's particular brand of sentimentalism.

The main advantage of Nichols' sentimental rules account is that it explains what he calls "core moral judgment." Core moral judgment has two main components: "a normative theory prohibiting harming others, and some affective mechanism that is activated by suffering in others" (Nichols 2004, 18). At its root, core moral judgment is the ability to distinguish between moral and conventional rules. Moral and conventional rules have in common that they are rules. To elaborate, they are normative principles that govern conduct and are abstract insofar as they apply to any agent that would perform a certain kind of behavior, even when variables may change.

For instance, consider the rule against a student talking in a classroom without first raising her hand. The rule applies to all students, in a wide variety a situations within the classroom. Therefore, understanding the rule requires the ability, given a classroom setting, to abstract and generalize from the situation. However, important differences hold between moral and conventional rules. Certain categories of inappropriate actions, such as hitting or pulling hair, seem different from other categories of inappropriate

[2] Cited in D'Arms and Jacobson 2000, 722.

[3] In more technical terms, D'Arms and Jacobson refer to the heart of neosentimentalism as the "Response Dependency Thesis:" "to think that X has some evaluative property φ is to think it appropriate to feel F in response to X."

actions, such as talking over everyone in the room or slurping one's soup. The first category contains morally impermissable actions, while the latter contains conventionally impermissible actions. Nichols surveys psychological experiments (including some of his own) that indicate that young children can make this distinction, but psychopaths cannot. The reason for this difference is that young children respond affectively to the suffering caused by pulling hair, but psychopaths do not. As Nichols' puts the point, "children respond with distress and concern to another's suffering. These responses seem to be diminished in psychopaths" (Nichols 2004, 17). Nichols cites Blair's research, which finds that even nonpsychopathic criminals make a significant moral/conventional distinction, whereas psychopaths do not (Blair 1995, 1–27).[4] When asked why one should not pull someone's hair, young children will say something like "because it will hurt them," whereas psychopaths will say something like "because it's not allowed" (Nichols 2004, 19).

Therefore, core moral judgment premises a distinction between moral and conventional rules. This distinction is rooted in an affective response to the suffering an action causes another person, and even young children can make this distinction. Furthermore, people with core moral judgment attribute other characteristics to moral rules they do not attribute to conventional rules. In particular, an action that violates a moral rule usually ranks higher on a scale measuring the seriousness of the violation, the generalizability of the violation (it is wrong in all cultures, not just the subject's own culture), and the independence of the violation from the presence of an authority (Nichols 2004, 22).

All these attributes tend to apply to actions that harm another person. The connection between the 1) affect and 2) seriousness, generalizability, and authority-independence combine to produce the distinction between moral and conventional rules. This distinction rests on whether or not the action involves an affective response. To break conventional rules does not normally activate the agent's affective mechanism. Thus, people see those deeds as wrong only because an authority disallows the action. However, actions that harm another person activate the agent's affective mechanism, and it seems that actions that cause harm and activate the affective mechanism closely tie to universal harm norms for most (nonpsychotic) agents.[5]

[4] Cited in Nichols 2004, 12.

[5] Even psychopaths have a normative theory prohibiting hitting for instance, but, as previously stated, they offer conventional justifications for their moral violations. The main difference between the psychopath and the nonpsychopath seems to be the activation of affect in response to another's suffering. It should be

To sum up, Nichols' sentimental rules account posits that core moral judgment entails the ability to distinguish between moral and conventional rules. Moral rules activate the agent's affective mechanism, so the agent sees the rules as serious, generalizable, and independent of authority. The final, crucial point is that "the affective mechanism responsive to suffering in others, in conjunction with information about harm norms, produces the nonconventional theory" that guides moral judgment (Nichols 2004, 27). So, the agent must have both an affective mechanism responsive to suffering in others and the ability to understand and apply abstract norms regarding harm. One should keep in mind the two main aspects of Nichols' sentimental rules account as examination turns to the movies and de Waal's evolutionary moral sentimentalism.

Section Two: The Evolution of the Moral Community in *The Planet of the Apes* Reboots

The overall narrative arc of the recent *Planet of the Apes* movies shows the development of the great apes into a human-like community where the members have human-like capacities for language, thought, and culture. The 2011 *Rise of the Planet of the Apes* tells the story of how the first chimp develops those human-like capacities and shares them with other great apes. In *Rise*, a common chimp develops human intelligence through exposure to an experimental Alzheimer's drug in utero, purported to stimulate neurogenesis, or the growth of new brain cells. The baby chimp is able to hold his own bottle at just a few days old, solves complex puzzles, tests at human IQ levels, rapidly picks up sign language, and adapts to life in a human home. The people caring for him name him Caesar.

In addition to improved motor skills and cognition, Caesar exhibits reflective consciousness and emotional intelligence sometimes thought of as exclusively human. While walking in the park with his adoptive father, Will Rodman, Caesar notices a dog on a leash. Caesar is also on a leash, and, after responding to the barking dog with a primal growl that reduces the poor canine to a whimper, Caesar signs to his father, "Am I a pet?" When his father replies, "No, you're not a pet," Caesar signs, "What is Caesar?" (Wyatt 2011). This moment of self-recognition reflects our more advanced understanding (compared to when the original film was made) of self-recognition in primates.[6] We now know that all great apes (except gorillas) and a few other animals recognize themselves in

noted that Nichols is not entirely clear on how he thinks these two elements are connected.

[6] For instance, see Povinelli et al. 1993, 347–372.

mirrors. Such recognition suggests that great apes understand themselves as individuals and potentially can distinguish their own mental states from the mental states of other great apes.

In a later scene, for example, Caesar witnesses an angry neighbor accost his human "grandfather," Charles Rodman, who suffers from Alzheimer's and has mistakenly entered, and damaged, the neighbor's car. Caesar perceives one of his caretakers to be in danger, cannot control his anger, and brutally, relentlessly attacks the neighbor in retaliation.[7] The assault leads to Caesar's confinement in a primate shelter. Caesar's human-like cognitive traits continue to develop. At first overpowered and bullied by the other apes, Caesar devises a plan to build a primate community with himself as the leader. He releases a large gorilla named Buck, who is kept in solitary confinement. Through this deed, Caesar gains Buck's trust and uses the gorilla's size to overpower, and thereby gain dominance over, the shelter's antagonistic Alpha male.

Caesar later gives the "Alpha" male a cookie, lets him out of his cage, and appoints him to distribute cookies to the others. To build community, Caesar then lets all the apes out of their cages and introduces them to a principle that will become one of the ape mottos: "ape together stronger." Caesar breaks free from containment and returns to his human home. Once there, he steals from the refrigerator the drug that has caused his extraordinary development. He then returns to the primate shelter and releases the drug, which exists as a gas, spreading it in a cloud to all the inhabitants. When they awaken the next morning, they too have increased cognitive capacities. Caesar then leads the apes in an escape from the primate center, through the trees of San Francisco, across the Golden Gate Bridge, and into the redwoods.

The second reboot, released in 2014 and titled *Dawn of the Planet of the Apes*, begins by showing the peaceful colony of apes Caesar and the others have founded. In the colony, the apes teach their children sign language, a written alphabet, and several abstract moral principles. For example, the apes live by the motto "ape not kill ape," an abstract, general moral principle that "unenhanced" apes cannot comprehend and therefore do not hold.

The end of the first movie and the beginning of the second also explain how the human race begins its devolution. The experimental drug that enhances nonhuman

[7] It is important to note that Caesar refrains from killing the man of his own accord. We know based on Chimp behavior in the wild as well as recent accounts of human raised chimps—such as the event with Travis in 2009—that many chimps Caesar's age would kill the man. For more information on Travis, see the following commentaries: de Waal 2009a and de Waal 2009b.

primate intelligence is deadly to humans. The drug infects a lab worker named Robert Franklin, who, while trying to let Will Rodman know he is ill, then inadvertently infects the same neighbor Caesar attacked. The infected neighbor is a pilot. The final scene of the first film shows the pilot going to the airport, boarding a flight, and rapidly spreading the virus across the globe.

The beginning of *Dawn* reveals the horrific loss of life that resulted from the "Simian flu."[8] Only 1 in every 500 humans, genetically immune to the virus, were spared its deadly effects. One group of human survivors has built a community in San Francisco, but the community is running out of power. So a small group travels into the redwoods, hoping to repair a dam that can provide their community with electricity. The humans and apes forge a fragile peace, but that peace breaks down after misunderstandings and betrayals of trust from apes and humans alike. By the end of *Dawn*, a war has begun between apes and humans.

Section Three: De Waal on Primate Social Behavior and the Building Blocks of Human Morality

Having briefly described the evolution of morality in the reboots, I now discuss recent research in primatology to illuminate the films. Fascinating about the reboots are the specific ways they reflect recent research in primatology and related fields, particularly the research of renowned primatologist Frans de Waal. In this section, I describe empathy and normativity, the two main poles of de Waal's evolutionary moral sentimentalism (EMS), and point out scenes in the reboots that reinforce key conclusions of his research.

De Waal is perhaps the foremost proponent of the continuity between primates and humans. In particular, he argues that nonhuman primate social behavior evidences the building blocks of human morality. De Waal argues that all the great primates (human and nonhuman) are fundamentally social, that feelings of empathy are at work in nonhuman primates, and therefore that both humans and primates are caring *and* violent, selfish *and* nurturing. Since de Waal goes to great lengths to argue that morality slowly has developed over the long span of human evolution, I refer to his account as an "evolutionary moral sentimentalism."

De Waal is a self-avowed moral sentimentalist. According to de Waal, moral sentimentalism "firmly anchors morality in the natural inclinations and desires of our

[8] This is not too far-fetched since the immune system of a Chimpanzee is much more robust than a human's immune system. That is why chimps are tragically the subjects are many medical experiments, often infecting them with diseases such as Hepatitis.

species" and emphasizes the role of the emotions in human morality (2006, 18). One of the primary philosophical advantages of sentimentalism generally is that sentimentalism provides a naturalistic account of moral emotions, reasoning, and judgment. By naturalistic account, I simply mean that sentimentalist accounts refer to nothing outside of the natural world (i.e. God or the soul), and look to evidence in the natural and human sciences to validate their position.[9] De Waal goes a step further by telling a story about how the roots of morality arise from nonhuman primate behavior.

The foundation of de Waal's argument for the continuity of primate behavior and human morality is his claim that human beings are "social to the core," a core largely shared with primates such as chimpanzees and bonobos (2006, 5). This shared social core generates the capacities for cooperation, reciprocity, fairness, self-control, and more. Crucial for this paper's argument, the social nature is the source of the two basic building blocks of the moral life: the capacity for reactive attitudes such as empathy and sympathy, and the capacity for "adherence to an ideal or standard" or what de Waal calls "natural normativity" (2014, 187). I briefly cover each of these in turn.

One of the key aspects of de Waal's EMS is the claim that morality is rooted in certain kinds of reactive attitudes, particularly the propensity to have the feeling of another agent involuntarily aroused in one's self. More specifically, de Waal stresses the importance of empathy for the moral life, or the ability deliberately to adopt the point of view of other agents, to see and feel things from their perspective (2006, 39). De Waal says that empathy "covers a wide-range of emotional linkage patterns, from the very simple and automatic to the highly sophisticated" (2006, 41). The very simple and the highly sophisticated both are observed in primates and are an essential part of human morality.

De Waal calls the most sophisticated form of empathy "attribution," or fully adopting another's perspective, referred to as theory of mind (or sometimes simply "mindreading"). To adopt that perspective, the agent must not only have the ability to look for reasons for the other's emotions, but also be able to understand the other agent's mental states, what the other believes, desires, and so forth. De Waal (2006, 26) more succinctly defines sympathy as follows: "an affective response that consists of feelings of sorrow or concern for a distressed or needy other (rather than the same emotion as the other person)." So, empathy is recognizing and feeling what the other agent is feeling, while sympathy is recognizing what the other is feeling, and feeling concern or distress

[9] A more technical definition is that ethical naturalism is the view that ethical or moral facts reduce to other natural facts, where a natural fact is something that is the subject matter of the natural sciences.

for the other. Empathy requires mindreading; sympathy does not.[10] But empathy and sympathy both have a vital element in common: they are *reactive* attitudes—responses to suffering or distress in another agent—and are motivating mental states.

One of de Waal's favorite examples of primate empathy is the chimp Kuni, who tried to help a bird fly by climbing to the top of the highest tree in the enclosure, wrapping her feet around the tree branch to leave her hands free, and then spreading out the bird's wings and launching the bird into flight (2006, 31). Kuni seemingly understood the difference between a chimp's needs and the needs of a bird, and responded appropriately. She could read the bird's "mind" and responded to its suffering. De Waal cites a number of similar examples that he takes as sufficient proof that primates can adopt another's viewpoint. Chimps respond to the pain they see in another chimp; a chimp helped another chimp who had fallen into a moat; chimps will protect comrades who are being attacked. All those responses require a basic understanding of the other's situation and emotional cues, and the ability purposefully to respond.

Biologists see in the reactive attitudes of empathy what they call "reciprocal altruism." De Waal defines reciprocal altruism as exchanged acts that are costly to the performer but beneficial to the recipient (2006, 13). Biologists believe such attitudes have evolved because "[e]volution favors animals that assist each other if by doing so they achieve long-term benefits of greater value than the benefits derived from going it alone and competing with others" (13). This is counterintuitive to our common-sense understanding of altruism: altruistic acts are precisely acts that bring *no benefit* to the agent. Yet, biologists refer to reciprocal altruism as altruism because any form of assistance toward another creature that costs the agent something and does not bring immediate positive results to the agent seems to run counter to one of the basic precepts of Darwinian evolution: natural selection (Sober and Wilson 1998, 25). Nevertheless, it is easier to see how tendencies to perform actions that will be reciprocated would be evolutionarily beneficial. It is more difficult to see how altruistic acts evolved that do not benefit the agent.

One sees many examples of unreciprocated altruism, from lowly parasites to the great apes. For example, the trematode parasite *Dicrocoelium dendriticum* spends the adult stages of its life cycle in the liver of cows and sheep, but, during the long process that it takes for the eggs to get from feces back to the liver, it spends its time in an ant. Of the fifty parasites that enter the ant, one migrates to the brain of the ant where it

[10] As addressed below, autistic children demonstrate the capacity for basic moral judgment and sympathy, but generally do not appear to have the capacity for empathy.

modifies the ant's behavior, causing the ant to spend more time on the tips of grass blades, where the ant is more likely to be eaten by a sheep or cow. The brain worm then dies, while the others go on to live as adult parasites in the host (Sober and Wilson 1998, 18). Despite the lack of intentionality of any mental states on the part of the parasite, the behavior certainly looks altruistic since the brain worm dies for the sake of its parasite comrades.

De Waal cites an example with apparent intentionality. A chimp named Krom notices that his aunt Jackie is trying to get water out of a large tire suspended from a pipe. The tire is pinned behind several other tires and thus Jackie cannot withdraw any water. After Jackie gives up, Krom begins to remove the tires one by one until he gets to the tire with water in it. He carefully removes it without spilling any water and carries the tire to Jackie. Jackie drinks her water, and Krom walks away without any display from either party. Two points are relevant. First, as in the Kuni story, Krom reads Jackie's mind and responds appropriately. Second, Krom receives no benefits for his actions; the altruism has no apparent reciprocity.

There are a number of interesting examples of reactive attitudes and reciprocal altruism in the *Planet of the Apes* reboots. Recall the earlier example from *Rise*. Caesar's human grandfather, suffering from Alzheimer's, mistakenly wrecks a neighbor's car, and the neighbor confronts him. Caesar is aware of his grandfather's condition. In an earlier scene, for example, he helps the old man use the correct end of his fork to eat eggs. Thus, when Caesar sees his grandfather being attacked, he leaps into action to protect him.

In *Dawn*, Caesar responds to the suffering of another agent, that of his wife, who gets sick after childbirth (he starts a family in the chimp colony, in the interim between the events of the two movies). However, one of the most moving examples is how he responds to the small band of human survivors that need access to a dam located within the territory of the ape colony. He agrees to help them despite protest from his fellow apes, including Koba, who points to scars on his body as evidence of mistreatment in the Gyn Sys laboratory, exclaiming "Human work!"

Caesar's experience with humans has been largely positive because of his primary caretakers. Indeed, in one scene, he refers to his human "father" as "a good man." Moreover, Caesar can see that the band of humans is desperate. Thus, he allows them access to the hydroelectric plant and even instructs apes to assist with the work. This assistance is not a cold, calculated move made in the interest of the ape community. Rather, it shows Caesar responding to the needs of others. It exemplifies recognition of other people's needs, a sense of their desperation, and a compassionate response.

However, the best example of positive reactive attitudes, of reciprocal altruism, happens at the end of *Rise*. As the apes are attempting to escape across the Golden Gate Bridge and into the redwood forest, a police helicopter opens fire on Caesar, believing that if they take out the leader, they can stop the revolt. Buck, the large gorilla Caesar had released from solitary confinement in the primate shelter, pushes Caesar out of the way of gunfire, takes the gunfire on himself, and, leaping into the helicopter, sacrifices his life to take down the assailants. While this act returns Caesar's kindness, it goes far beyond the original act. Buck gets nothing in return for giving his life for the sake of a friend.

Reactive attitudes such as empathy and sympathy constitute one of the basic building blocks of human morality, and de Waal has long emphasized their importance. More recently, he has begun to emphasize normativity as well. In doing so, de Waal responds to frequent criticism from philosophers that the reactive attitudes of animals are not intentional; rather, animals are wantons, creatures that follow whichever desire is strongest.[11] For instance, one may argue that evolution has hardwired reciprocal altruism into the great apes because helping conspecifics brings potential future benefits to the agent, thus increasing the agent's chances of survival. But simply acting in accord with one's strongest impulse is not moral; morality often requires conformity to a standard even when a desire conflicts with that standard. De Waal responds to this criticism by arguing that many animals conform to norms, often in ways that resemble human moral action. He argues that, at the most basic level, we see normativity in animal behavior when spiders repair webs or ants repair the nest (de Waal 2014, 187). But this normativity is reflected in much more important ways, such as in instances of fair distribution of rewards, acts of self-control, and reconciliation. I briefly review these examples.

Many studies have suggested that some animals respond negatively to the unjust distribution of rewards or goods. For example, in a now-infamous experiment, de Waal and colleague Brosnan had two capuchin monkeys perform a simple task for a reward. The first monkey performed the task, received a cucumber slice, and appeared satisfied with the reward. The second monkey performed the same task, but received a grape as his reward. The first monkey performed the task again, but, when he was again given a cucumber slice, he revolted, throwing the slice at the experimenter. This first monkey

[11] Both Kitcher and Korsgaard use this term—made popular in the philosophical literature by Harry Frankfurt (1971, 5–20) in his essay "Freedom of the Will and the Concept of a Person"—in their respective responses to de Waal in *Primates and Philosophers*.

protested each time, since his partner continued to get grapes for the same work while he got cucumbers (Brosnan and de Waal 2003, 297–299).[12]

De Waal (2014, 195) calls this "inequity aversion" (IA). The reaction is even stronger when the experiment couples one agent's reward with the punishment of a conspecific. In another famous study, rhesus monkeys could receive food by pulling on a lever, but doing so delivered shocks to a conspecific. The experimenters found that many monkeys would refuse to perform this task. The aversion was so great that one monkey refused to eat for five days, while another refused to eat for twelve days (Masserman et al. 1964, 584–585).[13] The monkey's sense of fair distribution of goods coupled with the reactive attitudes in response to the pain and suffering of conspecifics proved a great motivator.

De Waal rightly notes that the capuchin monkey experiment exemplifies disadvantageous IA. The agent negatively responds to the unjust distribution of goods that is disadvantageous to the agent. A higher level of fairness is advantageous IA: the aversion to the unequal distribution of goods that *favors* the agent. This appears to be a more uniquely human capacity. However, recent experiments have tested chimps in a version of the ultimatum game and appear to lend some evidence in favor of primate *advantageous* IA. In the now-classic experiment, a human subject (the proposer) is given a sum of money, for example, 10 dollars. The subject has a partner (the respondent) who knows how much money the subject received. The proposer gets to choose how much money she can keep and how much to give to the respondent. The motivator is that if the respondent accepts the offer, then both participants keep their share. However, if the respondent rejects the offer, then neither participant gets to keep any money.

People in Western cultures typically offer around 50% of the available amount as do people in most other cultures (Guth 1995, 329–344; Camerer and Loewenstein 2004, 3–52; Henrich et al. 2001, 73–78).[14] Surprisingly, in a simplified version of the ultimatum game designed for chimps and 3-to 5-year-old children, chimps tended to opt for an equal distribution instead of an unequal one (Proctor et al. 2013, 2070–2075).[15] De Waal takes this to suggest that chimps are also sensitive to unequal distributions of goods that favor the agent. However, neither chimps nor children distributed the goods equally in the absence of partner influence, suggesting a lack of autonomous moral agency.

[12] Cited in de Waal 2014, 195.

[13] Cited in de Waal 2006, 29.

[14] Cited in de Waal 2014, 197.

[15] Cited in de Waal 2014, 197.

Nevertheless, taken together, these studies suggest that nonhuman primates are sensitive to unequal distribution of goods, respond negatively, and take action to attempt to rectify the situation by bringing it back in line with a norm of fair distribution.

Recall the time in *Rise* when Caesar appoints the former "Alpha" male, Rocket, to distribute cookies to all the apes. Caesar begins by giving a cookie to Rocket, then instructs him to give one cookie to each ape until the cookies run out. Caesar could have easily kept all the cookies for himself, or handed them out preferentially. Instead, Caesar seems to recognize the importance of fair distribution for community building.[16]

Fair distribution is one way that nonhuman primates seem to conform to norms. A second way is self-control, particularly impulse control. One of the main charges against primate moral instincts by philosophers is that primates are wantons—creatures that always follow their strongest desire. Documenting impulse control would go a long way toward demonstrating that nonhuman primates can check their stronger impulses for an alternative though less strong desire. In what follows I describe several experiments suggesting that nonhuman primates have this ability.

In another now-classic experiment, children are given a marshmallow and are promised that they will get another marshmallow if they can refrain from eating the first. Children can hold out for several minutes, but so can monkeys and chimps (Mischel, Ebbesen, and Raskoff Zeiss 1972, 204–218; Logue 1988, 665–709; Beran et al. 1999, 119–127; Amici, Aureli, and Call 2008, 1415–1419).[17] As interesting as they are, these experiments demonstrate only the ability for participants to delay gratification for a short time to get a greater amount of the same gratification.

Other studies involving intentional self-distraction are more illuminating. For instance, Evans and Beran put a spin on the delayed gratification experiment: they offered chimps toys to play with while the chimps were offered a treat to see if the chimps would distract themselves. Again, the chimps knew that, if they refrained from eating the treat, they would get a greater reward. The chimps played with the toys and ignored the treat, allowing them to delay gratification for up to 18 minutes. As a control, the experimenters ran the experiment with the reward outside the enclosure, out of the reach of the chimp, so there was no temptation to consume the treat before it had accumulated (Evans and Beran 2007, 599–602). In this instance, the chimps did

[16] My thanks to Les Ballard for pointing this example out to me.

[17] For the experiments with children, see Mischel, Ebbesen, and Raskoff Zeiss 1972, 204–218; and Logue 1988, 665–709. For experiments with chimps and monkeys, see Beran et al. 1999, 119–127; and Amici, Aureli, and Call 2008, 1415–1419. Cited in de Waal 2014, 189.

not bother playing with the toys, indicating they had intentionally played with the toys in the previous experiment to distract themselves from the reward. Although this is still simply an instance of delaying gratification for a time in favor of greater gratification in the future, this shows that primates can intentionally distract themselves, one of the most basic instances of human self-control and emotion regulation and a necessary skill for deliberation and future planning.[18]

Nonhuman primates demonstrate impulse control when presented with greater positive outcomes, but they also can control their impulses when faced with negative outcomes. When several chimps all want to mate with the same female, often they sit around for hours grooming each other and calming themselves down rather than engage in a vicious battle for her. No one approaches the female until each male is sufficiently calm, and this behavior wards off a violent altercation (de Waal 2014, 194–195). Chimps do likewise when they are expecting food, which often can cause an altercation. Ostensibly warding off a fight, they will groom each other and engage in celebrations (195). These crucial examples show instances of impulse control when faced with the possibility of a negative outcome. Furthermore, the impulse is being controlled not simply for the sake of greater future gratification, but to avoid painful conflict and maintain communal harmony.

When a member of the human colony approaches Caesar and asks to be allowed to repair a dam that can provide the human colony with unlimited power, Caesar takes a night to deliberate. Most of the apes want to attack the human colony. Koba—Caesar's close confidant—fears that electricity will give the humans more power, making them more of a threat to the apes, and insists that the apes do not help the humans. Caesar is partially afraid that if he does not help the humans, they will attack. After Koba responds, "Let them attack," Maurice—another confidant—points out they do not know how many humans there are, or how many guns they have. That uncertainty does not change Koba's mind. However, Caesar wants above all to prevent a war because he knows that war risks all they have built: home, family, and future. Koba cannot control his impulses, but Caesar can. Because of his impulse control, Caesar is able to engage in future planning and goal-oriented deliberation.

[18] There is a large and growing body of literature documenting the human capacity for emotion regulation and its relationship to the agent's overall welfare. For recent studies, see Feinberg et al. 2012, 788–95; and Lai, Haidt, and Nosek 2014, 781–794. For broader overviews, see John and Gross 2004, 1301–33; and Beauregard 2007, 218–236.

The final example of conformity to norms to highlight is reconciliation. De Waal notes, “about thirty different primate species reconcile after fights, and that reconciliation is not limited to the primates. There is evidence for this mechanism in hyenas, dolphins, wolves, domestic goats, and so on” (2014, 192). After a conflict, chimps will groom and kiss each other, while bonobos will engage in sexual contact. Reconciliation often is seen in preventative form as well. For instance, when young chimps engage in playful wrestling bouts, a mother steps in and stops the bout at the first sign of distress. Her mediation keeps a conflict from breaking out. Some of the above examples on impulse control are also about conflict prevention.

Conflict resolution is another example of nonhuman primates curbing certain behavioral tendencies that would negatively affect the community. If conflicts can be peacefully resolved (or better yet prevented from occurring), then a certain standard of communal harmony can be maintained. The reboots also highlight conflict resolution. In one of the more powerful scenes in *Dawn*, Koba enters into the dam to find humans and apes working together to repair it. Recall that the Gyn Sys lab experimented on Koba, and he has many scars on his body that he refers to as “human work.” Appalled that the apes are helping the humans, he demands to see Caesar. As he is looking for Caesar, Koba pushes a human teenager to the ground. He cannot control his impulse to do violence.

When Caesar emerges, Koba asks why Caesar insists on helping the humans, declares that Caesar loves humans more than apes—loves humans even more than his own son. Caesar erupts in anger at this comment, and the two start to fight. Caesar gains the upper hand and nearly strangles Koba to death; but Caesar stops short, pronouncing between pursed lips, “ape not kill ape.” Koba stands up, assumes a bowing posture of submission, extends his hand, and asks for Caesar’s forgiveness in front of many other apes. After briefly considering Koba’s gesture, Caesar extends his hand, thus accepting the act of reconciliation. Caesar is able to return to himself from a violent immediacy, apply the ape motto, abstract from the situation to prefer the universal of forgiveness over the particularity of violence, and reconcile with Koba. Although the reconciliation is short lived, it highlights the primate capacity for reflection and abstract reasoning, forgiveness, and reconciliation. Even primates do not want to live in a constant state of violent upheaval, so they have developed tendencies and practices that help maintain communal harmony.

Section Four: The Impartiality of Moral Judgment

Thus far I have highlighted two main aspects of de Waal's EMS—reactive attitudes and normativity—and documented how the recent *Planet of the Apes* films imaginatively portray these aspects. I also have described a recent version of moral sentimentalism, Nichols' sentimental rules account. A clear connection exists between Nichols and de Waal's versions of moral sentimentalism. The heart of Nichols' account is the capacity for core moral judgment, or the ability to distinguish between moral and conventional rules. The capacity for core moral judgments rests on the marriage of an affective mechanism with the understanding of abstract norms regarding harm. De Waal has argued that nonhuman primates have the ability for reactive attitudes, particularly in response to the needs of suffering of others, and that nonhuman primates can adjust their behavior given certain goals or communal behavioral standards.

At first glance, the connection between de Waal's EMS and Nichols' sentimental rules account appears to strengthen de Waal's case for the continuity between primate social behavior and human morality. However, on closer investigation, it becomes clear that to understand and apply abstract moral rules, an ape must "possess the representational processes necessary for systematically reinterpreting first-order perceptual relations in terms of higher-order, role-governed relational structures. . .," what Povinelli and colleagues refer to as the "relational reinterpretation hypothesis" (Penn, Holyoak, and Povinelli 2008, 111). In other words, the application of moral norms requires the ability to abstract oneself from one's particular position and consider how a general principle—the common underlying explanation—may apply to any person in a different situation that has certain features in common—yet is perceptually disparate—from the current situation.

For instance, in experiments 9–14 described in Povinelli and Ballew's (2012, 138) *World without Weight*: *Perspectives on an Alien Mind*, a group of chimpanzees are presented with various weight sorting tasks. In experiment 9, the chimps are trained to sort the objects based on weight, and put the object in one bin if it is heavy and in the other bin if it is light (it should be noted that the difference in weight is typically 10-fold). If they get it right on the first try, they get a treat. No chimp tested could learn to do this in fewer than 400 trials, while some took up to 1562 trials, with a mean of 895 trials (97). Experiments 19–23 measured the impact of weight. In one variation, the chimps had to choose one of two balls and roll it down an incline. Only the heavy ball would push an apple through a hinged door toward the bottom of the incline. If the chimp chooses correctly, then she gets the apple. Again, this takes hundreds of trials for the chimp to learn. In both of these studies (as well as in many others conducted by

Povinelli and his team), the chimps' ability to sort based on weight drops to mere chance (186). Alternatively, experiment 30 tested the ability of 3–5 year old human children to sort and understand the impact of weight, and found that nearly 100% performed the tasks correctly. In fact, children often pass it on the first try "without assistance from the main experimenter" (255). Children demonstrate the ability to understand an abstract concept such as weight that chimps apparently lack.

From these and other experiments, Povinelli concludes that chimps do not understand the concept of weight, which requires the ability to "reinterpret observable objects and relationships in terms of unobservable variables" (2012, 26). A clear connection exists between concepts such as weight, and moral principles. Based on the growing evidence that nonhuman primates are incapable of this kind of abstract thinking, the kind of moral reasoning and action creatively depicted in the recent *Planet of the Apes* films is indeed imaginary. While conflict resolution among nonhuman primates is well documented, Caesar's motivation for restraining from killing Koba in the dam scene is based on his commitment to an abstract moral principle that an ape cannot hold due to its generalizability. Furthermore, although we know that nonhuman primates can delay gratification, delaying gratification in favor of long-term goals such as the good of the ape community and its progeny clearly outstrip ape intellectual capabilities.

To be fair, de Waal never claims that nonhuman primates are capable of human morality. Nevertheless, it often seems that he wants to hide this fact. However, de Waal hints at his understanding of human morality, claiming that human morality differs only *quantitatively* from primate social behavior. In other words, de Waal's evolutionary moral sentimentalism posits a "total gradualism" between primate social behavior and human morality.[19] However, de Waal's own understanding of the distinctiveness of human morality coupled with our growing knowledge of primate intellectual capacities highlights a gap between primate social behavior and human morality. This gap is highly problematic for de Waal's evolutionary moral sentimentalism, because de Waal argues that human morality evolves from primate social behavior, but he cannot provide an evolutionary story to explain how the second major prong of his own theory—normativity—evolved. To understand the nature of this gap between primate social behavior and human morality, it is useful to consider de Waal's own understanding of the human moral sense.

When discussing his understanding of the evolution of human morality, de Waal often quotes Darwin: "Any animal whatever, endowed with well-marked social instincts,

[19] This is Christine Korsgaard's (Korsgaard 2006, 104) term.

the parental and filial affections being here included, would inevitably acquire *a moral sense or conscience*, as soon as its intellectual powers had become as well developed, or nearly as well developed, as in man" (Darwin [1871] 1981, 71–72; quoted in de Waal 2006, 14). De Waal makes mostly clear what the social instincts are and how they can develop into more complex mental states. However, the nature of this human *moral sense* requires illumination. De Waal's discussion of the late-nineteenth-century philosopher and sociologist, Edward Westermarck, lends clarity. De Waal endorses Westermarck's distinction between reciprocal attitudes and moral emotions. Whereas reciprocal attitudes such as "gratitude and resentment directly concern one's interests," moral emotions are marked by their "disinterestedness, apparent impartiality, and flavour of generality" (Westermarck 1906, 738–739).[20]

The impartial, disinterested nature of the moral emotions may seem to put them at odds with the core of reciprocal attitudes, attitudes that require the basic ability to recognize suffering in another agent and have sympathy for that agent. Here de Waal's nod to moral philosopher and economist Adam Smith's "impartial spectator" is helpful. Smith (along with his friend David Hume) believes that human beings have the unique ability to expand local dispositions of kindness, sympathy, and reciprocity directed originally toward children, kin, and perhaps other members of one's in-group. The truly moral emotions or sympathies, however, "should be moved by what is 'useful and agreeable' to people (in general)," even when that general good conflicts with selfish or local interests (Kitcher 2006, 132). On a Smithian account, this transition involves reflecting on or mirroring the various judgments and perspectives and combining them into a genuine moral sentiment (133).

Smith famously stated, "We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it" ([1759] 1982, 204). This "impartial spectator" is the inner moral faculty by which we judge ourselves. Furthermore, "it is the peculiar office of those (moral) faculties . . . to judge, to bestow censure or applause on all the other principles of our nature" (273–274).[21] De Waal refers explicitly to Smith's notion of the "impartial spectator" and states that in this area of disinterestedness human emotions "seem to go radically further than other primates' [emotions]" (de Waal 2006, 20). Smithian moral approval requires distancing ourselves from our personal standpoints to obtain an impartial view of our own motives. Therefore,

[20] Cited in de Waal 2006, 20.

[21] Cited in Kauppinen 2014.

de Waal seems to endorse an account of human morality at the same time beyond the reach of primate cognitive capacity, yet rooted in basic features of primate social behavior.

De Waal's recent attempt to provide an account of "natural normativity," or the ways that nonhuman primates bring their behavior in line with certain standards, connects to one of the most fascinating, difficult aspects of Smith and Westermarck's moral sentimentalist account. This aspect is also especially relevant to the films. On Smith's account, the "impartial" nature of moral judgment begs the question: Since one agent cannot possibly know or take into account every other agent's perspective, how does the spectator arrive at her impartial judgment? Smith offers a hint when he notes that most of our moral judgments are based on general rules, which are themselves rooted in our emotional responses to particular cases (Kauppinen 2014, 16; Smith [1759] 1982, 387). Similarly, Westermarck (1906, 4) notes, "To name an act good or bad, ultimately implies that it is apt to give rise to an emotion of approval or disapproval in him who pronounces the judgment . . ." Furthermore, the agent makes the judgment on the account of an "accepted general rule" based on an "emotional sanction in his own mind" (6).[22]

Although it is not entirely clear how these general rules and emotional responses are connected, here is one way to understand the preceding comments. Imagine that an agent witnesses a morally salient action—an act of fraud against a conspecific, for instance—but that this action does not directly concern the witness. The witness is in a hurry and is not affected by the incident: he does not feel particularly bad for the victim nor does he know her. In fact, the witness may never see the victim again. But he believes it is wrong to deceive another person, and one of the main reasons he believes that is wrong is the pain that he feels when someone else deceives him. Therefore, even though he does not actually feel sympathy at the moment, he is motivated to help the victim because of a general principle he is committed to, namely, that "it is wrong to intentionally deceive another person." That principle derives from his own reactive attitudes toward those who deceived him in the past. Furthermore, he believes that most other people are also hurt when they are intentionally deceived. In this way, abstract rules or principles derive from, and connect to, common emotional responses, even when those emotions are not active in the agent performing the moral judgment.

This discussion of the impartial nature of moral judgment is meant to highlight both an advantage and a disadvantage of de Waal's EMS. De Waal's recent attempt to demonstrate that many nonhuman primates adjust their behavior because of norms suggests they may be capable of impartial moral judgment. They appear to recognize

[22] Cited in Kauppinen 2014.

suffering in others and often act to alleviate that suffering. They work to prevent harmful conflicts from arising that would negatively affect the community. They control their impulses not only to realize more advantageous personal outcomes, but also to maintain communal harmony. Taken together, these points enhance our understanding of the roots of human morality in primate behavior; however, the moral behaviors still fall short of anything resembling impartial moral judgment. De Waal himself notes this point, not only when he admits that disinterested moral emotions go far beyond the reciprocal attitudes of primates, but also when he discusses natural normativity. When comparing his understanding of natural normativity with impartial moral judgment, he states:

> Differences likely remain, however. Other primates do not seem to extend norms beyond their immediate social environment, and appear unworried about social relationships or situations that they do not directly participate in… One could argue that their behavior is normative in that it seeks certain outcomes, but that animals manage to do so *without normative judgment*. They may evaluate social behavior as successful or unsuccessful in furthering their goals, but not in terms of right or wrong. (de Waal 2014, 200)

The basic point here is that what de Waal calls "normative judgment" requires the ability to formulate and apply abstract moral principles across dissimilar situations. It requires the marriage of an affective mechanism activated by suffering and the ability to understand and apply moral principles. Recent research in primatology indicates that nonhuman primates cannot perform the abstraction and generalization needed to apply moral principles. Until de Waal can explain how human beings develop this cognitive capacity through the evolutionary process, his evolutionary moral sentimentalism contains a major lacuna.

Conclusion

De Waal's evolutionary moral sentimentalism comes with problems. Nevertheless, one gains a valuable perspective by using it as a lens through which to examine the recent *Planet of the Apes* reboots. The correlation between de Waal's evolutionary moral sentimentalism and the evolution of morality in the reboots is clear. Many of the necessary steps for transitioning from reactive emotions to the founding and sustaining of a moral community are seen in the two films. We observe, for example, Caesar's capacity for reactive, though local, attitudes, such as when he protects his grandfather from an angry neighbor. We then see Caesar expand his sympathies as he begins to have similar attitudes toward all his fellow great apes, for example, when he shares cookies equally among the entire primate shelter, or when he refuses to kill Koba because "ape not kill ape." Caesar develops truly moral notions, and, once his compatriots experience similar cognitive increases, they display expanded sympathies as well.[23] This allows the apes to establish a moral community that understands and applies abstract (disinterested) moral principles.

I have argued that recent research suggests that, while primates exhibit reactive attitudes, self-control, and other protomoral capacities, they are incapable of abstraction and disinterestedness. However, the movies offer an imaginary glimpse into how these truly "moral" capacities develop. Once an individual acquires this ability and forms a community of agents with the potential for disinterested moral emotions, it seems natural to foster those emotions in the group through education, specifically, the communication of certain abstract principles that reflect the group's sympathies.

However, this ability to form a community based on abstract principles that stem from impartial moral judgments also defines a boundary between primate social behavior and human morality. The boundary certainly is fluid, but it helps us to recognize the nature of truly *moral* emotions. Moral emotions require abstract reasoning and disinterestedness, and, until research proves otherwise, we have good reason to conclude that our planet lacks any apes with an abstract moral code.

[23] It is interesting to note, however, that Caesar's moral sympathies and his commitment to moral ideals is much greater than in his conspecifics. Clearly he has been treated kindly in the human home in which he was raised, and he has been morally educated in that environment. Comrades like Koba were not so fortunate, and their moral capacities reflect their upbringing.

References

Amici, F., F. Aureli, and J. Call. 2008. "Fission-Fusion Dynamics, Behavioral Flexibility, and Inhibitory Control in Primates." *Current Biology* 18 (18): 1415–1419.

Beauregard, M. 2007. "Mind Does Really Matter: Evidence from Neuroimaging Studies of Emotional Self-Regulation, Psychotherapy, and Placebo Effect." *Progress in Neurobiology* 81 (4): 218–236.

Beran, M. J., E. S. Savage-Rumbaugh, J. L. Pate, and D. M. Rumbaugh. 1999. "Delay of Gratification in Chimpanzees (Pan Troglodytes)." *Developmental Psychobiology* 34 (2): 119–127.

Blair, R. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57 (1): 1–29.

Brosnan, S. F., and F. B. M. de Waal. 2003. "Monkeys Reject Unequal Pay." *Nature* 425 (6955): 297–299.

Camerer, C. F., and G. Loewenstein. 2004. "Behavioral Economics: Past, Present, Future." In *Advances in Behavioral Economics*, edited by C. F. Camerer, G. Loewenstein, and M. Rabin, 3–52. Princeton, NJ: Princeton University Press.

Cox, Damian. 2006. "Agent-Based Theories of Right Action." *Ethical Theory and Moral Practice* 9 (5): 505–515.

D'Arms, Justin, and Daniel Jacobson. 2000. "Sentiment and Value." *Ethics-Chicago*-110 (4): 722–748.

Darwin, Charles. (1871) 1981. *The Descent of Man, and Selection in Relation to Sex*. Princeton, NJ: Princeton University Press.

De Waal, F. B. M. 2006. *Primates and Philosophers: How Morality Evolved*. Edited by Stephen Macedo, Josiah Ober, and Robert Wright. Princeton, NJ: Princeton University Press.

———. 2009a. "Another Chimp Bites the Dust." *Huffington Post*, February 17. http://www.huffingtonpost.com/frans-de-waal/another-chimp-bites-the-d_b_167768.html.

———. 2009b. "Chimp Attack: Missed Opportunities." *Huffington Post*, December 7. http://www.huffingtonpost.com/frans-de-waal/chimp-attack-missed-oppor_b_383078.html.

———. 2014. "Natural Normativity: The 'Is' and 'Ought' of Animal Behavior." *Behaviour* 151 (2–3): 185–204.

Evans, T. A., and M. J. Beran. 2007. "Chimpanzees Use Self-Distraction to Cope with Impulsivity." *Biology Letters* 3 (6): 599–602.

Feinberg, M., R. Willer, O. Antonenko, and O. P. John. 2012. "Liberating Reason from the Passions: Overriding Intuitionist Moral Judgments through Emotion Reappraisal." *Psychological Science* 23 (7): 788–795.

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

Guth, W. 1995. "On Ultimatum Bargaining Experiments: A Personal Review." *Journal of Economic Behavior & Organization* 27 (3): 329.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review* 91 (2): 73–78.

Hume, David. (1777) 1975. *Enquiry Concerning Human Understanding and Concerning the Principles of Morals*. 3rd Edition. Edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Clarendon Press.

John, Oliver P., and James J. Gross. 2004. "Healthy and Unhealthy Emotion Regulation: Personality Processes, Individual Differences, and Life Span Development." *Journal of Personality* 72 (6): 1301–1334.

Kauppinen, Antti, and Edward N. Zalta. 2014. "Moral Sentimentalism." In *The Stanford Encyclopedia of Philosophy*. Stanford University. http://plato.stanford.edu/archives/spr2014/entries/moral-sentimentalism/.

Kitcher, Philip. 2006. "Ethics and Evolution: How to Get Here from There." In *Primates and Philosophers: How Morality Evolved*, 120–139. Princeton, NJ: Princeton University Press.

Korsgaard, Christine. 2006. "Morality and the Distinctiveness of Human Action." In *Primates and Philosophers: How Morality Evolved*, 98–119. Princeton, NJ: Princeton University Press.

Lai, Calvin K., Jonathan Haidt, and Brian A. Nosek. 2014. "Moral Elevation Reduces Prejudice against Gay Men." *Cognition & Emotion* 28 (5): 781–794.

Logue, A. W. 1988. "Research on Self-Control: An Integrating Framework." *Behavioral and Brain Sciences* 11 (4): 665–679.

Masserman, Jules H., Stanley Wechkin, and William Terriss. 1964. "'Altruistic' Behavior in Rhesus Monkeys." *American Journal of Psychiatry* 121 (6): 584–585.

Mischel, W., E. B. Ebbesen, and A. Raskoff Zeiss. 1972. "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Psychology* 21 (2): 204–218.

Nichols, Shaun. 2004. *Sentimental Rules*: *On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

Penn, D. C., K. J. Holyoak, and D. J. Povinelli. 2008. "Darwin's Mistake: Explaining the Discontinuity between Human and Nonhuman Minds." *The Behavioral and Brain Sciences* 31 (2): 109–130.

Povinelli, Daniel J., and Nicholas G. Ballew. 2012. *World without Weight*: *Perspectives on an Alien Mind*. Oxford: Oxford University Press.

Povinelli, D. J., A. B. Rulf, K. R. Landau, and D. T. Bierschwale. 1993. "Self-Recognition in Chimpanzees (Pan Troglodytes): Distribution, Ontogeny, and Patterns of Emergence." *Journal of Comparative Psychology* (Washington, DC: 1983) 107 (4): 347–372.

Proctor, D., R. A. Williamson, F. B. M. de Waal, and S. F. Brosnan. 2013. "Chimpanzees Play the Ultimatum Game." *Proceedings of the National Academy of Sciences USA* 110 (6): 2070–2075.

Reeves, Matt. 2014. *Dawn of the Planet of the Apes*. Blu-Ray. Beverly Hills, CA: Twentieth Century Fox Home Entertainment.

Smith, Adam, D. D. Raphael, and A. L. Macfie. (1759) 1982. *The Theory of Moral Sentiments*. Indianapolis: Liberty Classics.

Sober, Elliot, and David Sloan Wilson. 1998. *Unto Others*: *The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

"Top-US-Grossing Feature Films." *IMDB.com*. Accessed June 10, 2015. http://www.imdb.com/search/title?sort=boxoffice_gross_us&title_type=feature&year=1968

Westermarck, Edward. 1906. *The Origin and Development of the Moral Ideas*. London: Macmillan and Co.

Wyatt, Rupert. 2011. *Rise of the Planet of the Apes*. Blu-Ray. Beverly Hills, CA: Twentieth Century Fox Home Entertainment.

Journal of
Cognition
and
Neuroethics

Simulating the Informational Substance of Human Reality in *Queen City Jazz*

Susan V. H. Castro
Wichita State University

Biography

Susan Castro is an Assistant Professor of Philosophy at Wichita State University. After completing her undergraduate degree in math and computer science and working on the Human Genome project at UCLA, she became interested in philosophy and ethics of science. Her current work is Kant-centered but interdisciplinary, focusing on the value of cognizing as if, ranging from the roles of imagination to the problematic mode of cognition in moral conduct and artistic activity.

Acknowledgements

Thanks to Winston Salser for his mentorship and support. Without his mentorship I would never have considered graduate school in philosophy, much less had the gumption to attempt interdisciplinary work in cognition.

Publication Details

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

Citation

Castro, Susan. 2015. “Simulating the Informational Substance of Human Reality in *Queen City Jazz*.” *Journal of Cognition and Neuroethics* 3 (3): 27–65.

Simulating the Informational Substance of Human Reality in *Queen City Jazz*

Susan V. H. Castro

Abstract

Queen City Jazz is a 1994 somewhat post-apocalyptic, somewhat post-human novel in which Kathleen Ann Goonan explores the beautiful and terrifying potential of the combination of unlimited nanotechnology with "an unscrupulous philosophy." The unscrupulous philosophy within the narrative belongs to the nanoarchitect, Durancy, who imposes his own substantive conception of the good on a society that cannot consent. A second, more basic, unscrupulous philosophy structures the world in which *Queen City Jazz* takes place and underwrites the conditions that enable Durancy to do so. The first half of this paper outlines this philosophical structure and provides a metanarrative highlighting some of the most innovative and distinctive features of the work, for example the idea that the vestigial human pheromone system could be transformed into a powerful means of communication, as powerful an experience machine as any art form has ever been. The second section of this paper addresses the problem of how idea driven science fiction could function as an epistemic tool and what we might learn from *Queen City Jazz* by treating it as a thought experiment. I argue that the extended narrative of idea driven science fiction novels can ground an exploratory simulation in episodic cognition that paradigmatically serves as a rich context for public reflection and discussion concerning how we ought to move forward in science and society. By episodically immersing readers in a dystopic future, thus engaging readers in an affect-rich first person simulation of that possible future, *Queen City Jazz* challenges readers to diagnose what precisely has gone wrong in the Queen City. The final section addresses what we can learn from the experiment, assuming that it is well designed. I argue that it *shows* us the intrinsic value of work, and this has real implications for our technological ends. One of the scruples missing from Durancy's philosophy is that humans need, thus ought, to work.

Keywords

Episodic Foresight, Constructivism, Thought Experiment, Pheromones

Queen City Jazz is a 1994 somewhat post-apocalyptic, somewhat post-human novel in which Kathleen Ann Goonan explores the beautiful and terrifying potential of the combination of unlimited nanotechnology with "an unscrupulous philosophy." The unscrupulous philosophy within the narrative belongs to the nanoarchitect, Durancy, who imposes his own substantive conception of the good on a society that cannot consent. A second, more basic, unscrupulous philosophy structures the world in which *Queen City Jazz* takes place and underwrites the conditions that enable Durancy to do so. The first half of this paper outlines the basic philosophical structure of the novel and provides a metanarrative highlighting some of the most innovative and distinctive features of the work, for example the idea that the vestigial human pheromone system could be transformed into a powerful means of communication, as powerful an experience machine as any art form has ever been. The second section of this paper addresses the problem of how idea driven science fiction could function as an epistemic tool and what we might learn from *Queen City Jazz* by treating it as a thought experiment. I argue that the extended narrative of idea driven science fiction novels can ground an exploratory simulation in episodic cognition that paradigmatically serves as a rich context for public reflection and discussion concerning how we ought to move forward in science and society. By episodically immersing readers in a dystopic future, thus engaging readers in an affect-rich first person simulation of that possible future, *Queen City Jazz* challenges readers to diagnose what precisely has gone wrong in the Queen City. The final section addresses what we can learn from the experiment, assuming that it is well designed. I argue that it *shows* us the intrinsic value of work, and this has real implications for our technological ends. One of the scruples missing from Durancy's philosophy is that humans need, thus ought, to work.

To better frame the problem, consider that freeing us of the burden of work has been one of the central advertised goals of technological development. If work has an intrinsic value for humans, for example because work is essential to freedom, then the elimination of work would be a dystopic ideal. The utopian goal of technology would instead be to free us of the burden without freeing us of work. To put the argument in mundane philosophical terms,

1. Freedom *from* is not an end in itself; its value is derived from enabling our freedom *to*. Idleness, passivity, and rest belong between projects. They cannot constitute a life. Aiming for negation negates us. It is the exercise of our powers that give us life, that make us live. Humans need to work.

2. Our ultimate positive aims are not given to us. We must make them for ourselves.

3. The goal of technology should thus not be to free us from work; it should be to free us to work. More specifically the goal of technology should be to free us from the constant burdens of natural necessity, thereby enabling us to construct for ourselves the plans and pursuits that give our lives meaning and value.

4. The artist, the athlete, and the scholar are archetypes of the lives we might lead if unburdened from natural necessity. Surely there are others we have yet to

> discover. No one should be left with nothing to do. We will need to try things out and experiment in thought and deed to realize meaningful ways of life for everyone. This is the work for which technology should free us.

Scholars who find the argument above compelling are likely to have already been softened up by Aristotle, Kant, Marx, or other intellectuals. They already believe, by some description or other, that life is purposive activity and humans are purpose-originators. Artists and athletes, too, are already members of the pro-work choir. The argument is not likely to be so persuasive, or even to get any traction, with those who have an entrenched belief that work is life-sapping exogenous toil, that exercise is an excruciating expenditure of the mind or body, or that effort is a natural evil. The experience of work as an alien torturous constraint is a difficult obstacle to overcome in attempting to convince someone that work is a human need, thus assertoric arguments like the one above cannot get off the ground as long as our fantasies of lives of leisure remain untested. Neither a more elegant assertoric argument nor a plethora of empirical support will help here.

It may be tempting to write off those who deny the first premise by positing that their negative experiences of work have generated an irrational psychological bias that is resistant to counterevidence, but contemporary epistemology offers a broader base of resources for understanding experiential evidence and its role in value judgments. Third person testimony is no match for first person experience, and it should not be, when it comes to making the value judgments by which we live. The kind of evidence that would best support the judgment that humans need to work is first person experience. Given that few if any of us can try out what it's really like to not work, it is fortunate that first person experience can be acquired through simulation. Thought experiments like Gyges' ring and Black and White Mary have been used for thousands of years to *show*, what cannot be said or told (Brown and Fehige 2014; Wittgenstein 1922). Novels like Kathleen Ann Goonan's *Queen City Jazz* can do so as well. As Goonan herself cleverly puts it:

> "You're rude and irritating and obfuscating." … "You could at least tell me what's going on in this place," she said.
>
> "You would think so, wouldn't you?" he asked. "That does sound reasonable, on the face of it"… "I wish that I could just *tell* you these things. But that way doesn't work – see?" (Goonan 1994, 210-11)

In this paper I both demonstrate and explain how a particular idea driven science fiction novel, *Queen City Jazz*, tests the life of leisure fantasy and brings us to confront the fundamental question "What (work) would we do?" if technology really freed us. The first half of this paper outlines the philosophical structure of *Queen City Jazz*, i.e. the system of ideas that set its parameters, and provides a metanarrative for how the novel proceeds. This provides a description of method for the experiment and allows readers to *do* enough of the experiment to *see* at least some of what it might show. The second section addresses theoretical problems concerning how science fiction novels could work epistemically, as extended thought experiments or experience machines. The upshot is that by recruiting our affect-rich first person episodic powers of simulation[1] they allow us to have counterfactual experiences that can counter our real experiences, e.g. of work, and consequently to make judgments informed by this extension of experience, as embodied, vulnerable, sentient beings who identify with the characters and critically reflect from the first person. The final section returns to *Queen City Jazz* to explain how this particular thought experiment supports the argument for work by viscerally communicating to readers that we need to *do* work of our own making, to *exercise* freedom, before closing with the suggestion that the aims of mental health research should respect the value of work in human life.

Queen City Jazz

Queen City Jazz is a 1994 somewhat post-apocalyptic, somewhat post-human novel in which Kathleen Ann Goonan explores the beautiful and terrifying potential of the combination of unlimited nanotechnology with "an unscrupulous philosophy." The unscrupulous philosophy within the narrative belongs to the nanoarchitect, Durancy, who imposes his own substantive conception of the good on a society that cannot consent because the vestigial human pheromone system has been transformed into a powerful means of communication, an experience machine as powerful as any art form has ever been. At this level, the novel is a quasi-aporetic exploration of what *we* would be like, so changed, both as individuals and as social beings.

[1] The episodic simulation theory I advocate in this paper is not necessarily in competition or conflict with others recently offered. For example Thagard's theory of intuition, which involves encoding representation, neural binding, and interactive competition, is a very low level theory that posits the neurological mechanisms by which *recognition* is possible. Given that recognition is presumably a component or precursor of relevance determination, his theory is an important step towards understanding how a thought experiment could work in the higher level processes (self-interrogation, judgments of relevance, and other skills) that Camilleri argues are necessary to execute a thought experiment *well* (Camilleri 2014). I take the position I advocate in this paper to be sufficiently generic to be compatible with both Thagard and Camilleri.

A second, more basic, unscrupulous philosophy structures the world in which *Queen City Jazz* takes place and underwrites the conditions that enable Durancy to do so. The most basic organizing principle of the novel is the metaphysical premise that *everything is information*, where information is both noun and verb, both the form a thing has at a given time and the processes of taking in or taking on new forms (Goonan 1994, 227). Because "information" is not essentially representative much less factive in *Queen City Jazz*, the principle that everything is information divorces value from actuality. This leaves nanotechnology with a blank evaluative slate and unlimited potential to rebuild the world from its elements up through human cognition and sociality. The most striking implications of this premise arise from an apparently rational choice to enhance the human limbic system, which both abrogates subsequent consent and enables humans to live Durancy's ideal of the artistic life. The dystopian aspects of the outcome then explicitly result from neglect of a natural law, namely that *any form of information transmission suffers loss*. The novel is thus ultimately structured by the following principles[2], which serve as parameters for the thought experiment that reveals a fatal flaw in Durancy's design. Rather than enabling us and freeing us, tampering with the limbic system is likely to generate addiction, dependence, and misery.

1. Everything is information.

2. Learning is the acquisition of information.

3. Nanoeducation is virtually limitless.

4. The human limbic system constitutively informs human experience and higher cognition.

5. The human olfactory/vomeronasal system could learn to become a high intensity, broad bandwidth informational channel.

6. Any form of information transmission suffers loss.

[2] I describe this as a set of principles rather than an argument because the logic of *Queen City Jazz* is somewhat inconclusive or aporetic, as any argument from an extended narrative thought experiment can be expected to be. It may be useful to think of the argument as an elenchus, as in a Socratic dialogue. Its purpose is to engage the reader in a cognitive process that one hopes will lead to new insight through the participant's own cognitive power.

7. Addiction is a result of information loss.

(Aporesis)

To understand the logic of *Queen City Jazz* it is necessary to take the first principle quite literally. At the elemental level, lead is *informed* by its atomic structure as well as by temperature, pressure, and other exogenous sources of information. At the human level, we are informed by our genetics, epigenetics, sensory experience, and memory; by the communication of an idea, by the ingestion of nutrition, and by inoculation against pathogens. Given that the acquisition of information is *learning*, all these mechanisms of information - of becoming informed - are forms of education. Education is thus literally transformative in *Queen City Jazz*.[3] In the novel as well as in reality, education is also often involuntary, irreversible, and dangerous. Given that nanotechnology is an artificial mechanism of information, one that need not respect natural kinds, nanoeducation is a nearly limitless learning mechanism in *Queen City Jazz*. Lead can learn to be gold. A train station can learn to repair itself after a bombing. You could wake up one fine morning to find a tiny blinking "*n*" on the ankle of your work boots. Who knows what they might have learned to do?

Because "information" must be construed so broadly in order to encompass literally everything, it necessarily encompasses what we would pre-theoretically call misinformation, falsehood, delusion, and corruption. Just as growth, development, and healing are processes of information – changes in form – so too are disease, consumption, trauma, and death. The principle that everything is information is entirely indiscriminate. Discrimination between good and bad forms of information would require an independent scruple. By constructing a world with this gap between fact and value, i.e. between the actual forms things take and the forms things ought to take, *Queen City Jazz* forces the reader to recognize both the indispensability of evaluation in and for human life, and how much our actual evaluations rely on unquestioned internalized norms (Stanley 2014; Dodd and Stern-Gillet 1995; Longino 1987).[4] The lesson for us all,

[3] *Queen City Jazz* provides a somewhat ironic perspective on the transformative education model according to which education is a lifelong project of transforming the person, not merely equipping the student with facts and skill sets (Boyd and Meyers 1988; Taylor and Cranton 2012). Personal transformation is a particularly important and elusive goal in ethics education and for building "sustainable societies."

[4] Largely driven by race and gender studies, there has recently been a sea shift in the incoming generation of philosophers towards a consensus that value free science is neither possible nor ideal. The myriad ways in which scientific practice and scientific theory are unavoidably value-laden was an unmistakably prominent

including neuroscientists, is that we cannot afford unscrupulous philosophies. If we are to build our world well, we must consider with extreme care what constitutes good science. If we are to make minds healthy, we must distinguish between good health and poor health, and we must do so without defaulting to the naturalistic fallacy that whatever is natural or normal ought to be our norm. *Queen City Jazz* makes vivid and visceral the tension between our need to divorce *ought* from *is* and our competing need to somehow ground what ought to be in natural norms.

The backstory of *Queen City Jazz* is that when an astronomical event early in the twentieth century makes the end of all energy-based broadcast communications imminent (Goonan 1994, 318), the human imperative to maintain the informational connections that define us drives a movement to preserve what we are by transforming our bodies.

> The whole world became dependent on this new system that they thought up, the [Enlivened] Flower Cities. Nan. Changing the human body itself to receive [chemical] messages so that everything else would change… (Goonan 1994, 162)

In the new system, everyone would be free from disease, free from material want, and even free from labor because the pervasive nanotechnic enlivenment of inorganic materials empowered inhabitants to literally realize whatever they could imagine.

> If you wanted a piece of 'wood' in a Flower City…You just went to your computer and ordered the substance, which was just like wood, every molecule, except that it wasn't really wood…yet there was no difference except in how it came into being…But it could also look like one thing but be different… (Goonan 1994, 152)

The substance you ordered could be inflammable wood, unbreakable wood. Not even the natural kinds of chemistry and physics are impervious to nan in *Queen City Jazz*. The Flower Cities would be our ultimate tool: self-sustaining, self-healing solar powered complex living beings that do for us whatever we do not wish to do for ourselves.

At first the Conversion to Enlivened Flower Cities surpassed expectations. According to the pamphlet welcoming newcomers to Cincinnati after its Conversion,

theme at the 2014 Philosophy of Science Association meeting, though the indispensability of ethical and social values to the content of science has often not been evident in prominent works in analytic philosophy of science (Bechtel 1988).

> …[They enjoyed] a standard of living unparalleled since the beginning of time. Communication has not only been restored, it is conducted at a faster rate and with both a greater accuracy and a wider emotional bandwidth than ever before… (Goonan 1994, 183)

The pamphlet explains how information is stored in DNA and bacteria, how the "pheromone breakthrough" enabled chemical broadcasting, and how its giant Bees and Flowers manage the information of the Enlivened city (Goonan 1994, 183).

> …the things nan could do were as dangerous as they were beautiful, and that was what made it so fascinating...the very shape of matter could be shifted and changed and used, almost as easily, once it was all set up, as just *thinking* about it…that was so glorious. (Goonan 1994, 79)

No more tedious trial and error in the laboratory. Our new metapheromonal interface to the nan assemblers would program them to execute our desires without all the fuss and bother.

The infectious idea that got it all started was articulated in 1984 by Eric Drexler in *Engines of Creation*:

> In physical terms, it is clear enough why advanced assemblers will be able to do more than existing protein machines. They will be programmable like ribosomes, but they will be able to use a wider range of tools than all the enzymes in a cell put together. (Goonan 1994, 149)

We can easily imagine how it might have happened, given that the nano-infrastructure for radically transforming humans is already being built.[5] Ribosomes and other natural molecular machines could be customized to denature and refold proteins as a kind of prion dialysis, or to sequester and release neurotransmitters to optimize brain chemistry (Sanbonmatsu 2012; Doyle *et al* 2013; Südhof 2013). Gene therapy packages will someday not merely splice DNA but also synthesize and deliver whatever additional infrastructure is needed to effect a new phenoform, from transcription factors to missing

[5] Current work in molecular machines already supports the eventual realization of almost everything mentioned here, *if* targeting problems become sufficiently tractable, though of course it is unlikely that the use of nanotechnology will ever be as easy as just thinking about it. For general overview of molecular machines see Coskun *et al* 2012 and Frank 2011.

elements of signaling cascades. Whole new signaling cascades with functions of our choosing might be engineered de novo by artificial molecular factories. DNA methylation machines and RNA nano-factories could reverse adverse epigenetic effects and implement epigenetic enhancements.[6] Advances in dynamic field theory[7] might allow us to cure brain development disorders and lesions through a growth factor orchestration of neurogenesis, differentiation, and selective connectivity to redevelop microstructure and macrostructure.[8] Brain plasticity could be finely managed to facilitate specialized learning, e.g. how to use a new artificial limb or to learn the languages of terrestrial alien beings like bees. In our scientific dreams we will soon be able to make of ourselves what we will.

Most perspicuously to Goonan's plot, in response to the energy-based broadcasting crisis our vestigial pheromone systems[9] could be redesigned to link in to the Flower City chemical information networks and link up to our conscious awareness.

> Once a human is genetically programmed, their own personally generated pheromones are re-assembled into metapheromonal[10] packages capable of precisely echoing the most complex thought humanity can achieve. Or the most simple. That package passes through the membrane at a touch, to be carried upward to the Flower

[6] The epigenetics of cognition are being researched on several fronts (Zannas and West 2014; VanHook 2015; Lattal and Wood 2013; Masri and Sassone-Corsi 2013; Molfese 2011).

[7] Dynamic Field Theory lends itself well to the sort of informational theory of experience underlying *Queen City Jazz*, particularly in that extension of the classic dynamic fields like vision to include computer networks and social fields would support hints of the extended mind hypothesis in the novel, e.g. dead reckoning scaffolded on semiochemical awareness of solar road location and hints at communal cognition in the Flower Cities (Sandamirskaya *et al* 2013; and Gallagher 2013).

[8] The basic components of very complex orchestrations, e.g. scaffolding, sequencing, and integration, are being developed (De Bo *et al* 2014; Du *et al* 2012; Yan *et al* 2013).

[9] Pheromones, and semiochemicals more generally, have been extensively studied in insects, especially bees and ants, and to a lesser extent in plants and other organisms. Though oxytocin and other putative human pheromones or semiochemicals are currently being studied, it is not yet entirely clear whether there is an operative human pheromone system. Our vomeronasal system is clearly vestigial, but the extant infrastructure is sufficient to pose the question: If we could restore, activate, or enhance this system, how would cognition be affected? (See Dölen and Malenka 2014; Carter 2014; Doty 2014).

[10] The "meta" in "metapheromone" indicates a higher order of informational complexity or content. There is little evidence in the novel that the metapheromonal system she envisioned is metacognitively present to consciousness or that metapheromones have any phenomenal quality, e.g. qualities akin to redness or saltiness.

> via bacterial DNA. There, in a form modeled on pollen, it can be collected and taken wherever needed, deposited and carried downward to the exact target room, and either be directly absorbed by the target or translated to any sort of tangible display. (Goonan 1994, 217)

The idea is that instead of conveying our thoughts through words or works of art, we could chemically encode and transmit them through the touch of our palms to the Flower network. Some of these thoughts might be programs for the city to execute. Others might be messages for the Bees to deliver directly to other human beings. As we will see, one of the plot-driving details of this system is that rather than collecting metapheromonal pollen on their legs like ordinary bees, the Bees of a Flower City would embody our thoughts.

> Human limbic tissue is integrated into the brain structure of every Bee…This gives them the necessary incentive for the work they must do, and binds them to the city, to humans. In this way they can carry complex emotional information… (Goonan 1994, 274)

To restore broadcasting capacity using this new chemical platform, the Flowers could diffuse semiochemicals through the air from atop their buildings, but of course *we* would have to be educated to receive them. Our olfactory and vestigial vomeronasal systems would have to be taught to become a high intensity, broad bandwidth informational channel. Nanoeducation would have to create a direct channel to the limbic system[11], which constitutively informs human experience and higher cognition. Scent has the power to evoke memory, mood and emotion. Pheromones can evoke desire and aversion. Nan could give us an olfactory capacity any dog might envy, and a vomeronasal capacity beyond peer. Most importantly, if the prefrontal cortex can recruit the visual cortex to envision in imagination[12], a direct channel of communication to an "educated" limbic system could recruit (or hijack) the entire brain, thus the entire body, to *experience*. The pheromone breakthrough thus could not have been a merely somatic breakthrough. It would have unavoidably transformed *us,* and made us receptive to exogenous emotion and volition, vulnerable to a plague of education gone wild (Goonan 1994, 74).

[11] For an overview of the limbic system and its relevance, see Catani *et al* 2013.

[12] For neural recruitment in imaginative activities see for example Buchanan *et al* 2014, Slotnick *et al* 2012, Gandevia *et al* 1997.

The Flower City utopia was of course short-lived. The assemblers improvised like a jazz band, spreading uncontrollably in wild Surges of material transformation and plagues of informational infection, many of which were literally dead ends.

> …people were dropping like flies…The plague takes everybody different. Makes them *learn*, you know? Does something to the brain. Supposed to enhance things. Make everyone superhuman… Only problem is, it got out and spread like wildfire before it was perfected. In fact, a lot of nanoplagues did…Apparently there's a lot more to being human than meets the eye. (Goonan 1994, 64-65)

Those who contracted the plague and survived were made *strange,* pheromonally driven by new imperatives and immersed in a reality constituted by new senses that were cognitively scaffolded on the immense nan-pervaded material world. Victims of the Norleans plague, for example, obsessively build Huck Finn rafts and sing their way downriver to New Orleans, or more often to their deaths in the rapids.

Queen City Jazz opens in the post-Surge world in which nan has become endemic, on the verge of a radical transformation in the life of the young protagonist, Verity. Though she has been raised in an isolated neoShaker community that shuns nan and deeply fears infection, Verity has nubs behind her ears. These are "proof of some sort of tampering; tampering which might infect the Shakers in some unknown way or even kill them" (Goonan 1994, 5). The Shakers are aware of her nubs but she is otherwise able to pass as an uninfected natural human. She hasn't told them that she and her dog Cairo share pictures in their minds. They are also unaware that she is an unnatural dead reckoner. Her cognitive maps are scaffolded on the nan-built solar road system (Goonan 1994, 86). Most importantly they do not know that she is annually Called to the Dayton library to be programmed.

When Verity returns from a programming session at the library, she always has new Dances. Following a preliminary state of seizure, the dance is executed in a cognitive state reminiscent of artistic or divine transport.

> Verity felt the Great Blessing echo through her body, unfolding like a flower of light which drew brilliance from the air around her straight into her body, and then it gathered into the center of her bones, concentrated, bright, and rushed upward through her spine until it flowered somewhere above the top of her head.

> She began to jerk… about five minutes, and the light within her grew more bold and warm, and if she opened her eyes she knew that all would be bathed in the light, and when she looked at the faces of those around her it would be as if this had all happened a million times before.
>
> The light pulled her from her seat … as she felt the Dance form and then propel her.
>
> She whirled … and began a complicated, repetitive step.
>
> She heard Blaze begin to play once more, as if from far away, a melody which hummed like a swarm of bees … and she heard the shuffling steps of others as, one by one, they joined her. …[T]hey scattered, reformed, swirled, and finally stopped, all in the same moment, as if they had practiced but they had not ….
>
> They had found that they were of one mind about her Dances. Sometimes, during Meeting, one of them would rise, and dance a few steps, and the others, remembering exactly, would join in, and for a time they would be part of something larger. (Goonan 1994, 18-19)

By packaging the Dance as at once seizure, enlightenment, and artistic performance, Goonan provokes us to consider the extent to which these phenomena are truly disparate. By positing kinesis as a mechanism for activating nanoprogramming, she further provokes us to consider the extent to which these phenomena are fundamentally somatic. Kinesis is a mechanism of communication in at least two ways. Interpersonally, body language communicates affect, both symbolically (e.g. skipping is a kinetic sign of joy) and sympathetically through mirroring. Kinesis is also an intrapersonal mechanism of communication through sensorimotor feedback loops that aid proprioception and fine motor control as well as cognitive orientation and equilibrium. As Goonan describes it, movement provides chemical information to the brain, which can then feed back to activate precise biochemical pathways that activate latent information (Goonan 1994, 276). Through this process one might learn to play the piano, or program nanoassemblers, perhaps without being metacognitively aware that one is learning or that one has learned. Nanoeducation could artificially produce *know how*, educating everything from declarative and episodic memory to muscle memory. When an entire community is primed or preprogrammed for reception, a Dance might teach them anything.

The pretense that Verity's community is an uninfected, natural human community is obviously just that – a pretense. They cling tightly to the illusion that they can control what informs them, though it is painfully obvious that they cannot. Other members of the community have their idiosyncrasies. For example, Tai Tai keeps a journal filled with "brackets, dots, numbers, letters, all jumbled together crazy and tight" and Blaze has a special interest in trains, one so distinctive and inexplicable that it seems artificial and exogenous (Goonan 1994, 14). After catching a cold one day Blaze inexplicably *knows how* to play Scott Joplin's "The Chrysanthemum," and he *knows that* a man named Scott Joplin wrote it. Perhaps Tai Tai and Blaze have caught interesting colds before. Perhaps they all have.

The nan contagion is not limited to the individual level, either. The founders built the entire community and wrote its Scriptures in a plague-driven fever.

> What do you think?...That Mother Ann [the original founder of the Shakers] appeared on the edge of Bear Creek in a pioneer dress with angel wings and handed that stuff over? Hell, no! Ma put it together in a frenzy, one fine summer just after she was infected...It was ecstasy, all right. (Goonan 1994, 66)
>
> …they raved, they *built*…with winches and saws. Built the barns, this house, the library…They were mad, Absolutely mad. Took them two years… Took a hell of a lot of energy *not* to give in and raft down the river…My mother was crazy. She believed that she was the manifestation of Mother Ann, sent to purify the human race. When people were dropping like flies… (Goonan 1994, 64)

The original founder of the Shakers, Mother Ann had envisioned a society in which men and women would be equal from the start, a celibate society that would effectively return human society to the Garden of Eden. Ma envisioned a society in which men and women could be natural from the start, a technologically celibate society that would return humans to the pre-Surge world. Whereas as Mother Ann had been raised in a sexist society, Ma had been infected by nan. Both sought escape for themselves and their community.

Despite her nubs and her special abilities, Verity has been brought up to fear contamination by outsiders who might carry plague. Yet when she eventually meets a survivor of the Norleans plague, the woman claims to be quite well.

> Getting the plague is the most wonderful thing that could ever happen to you. Plague!…That's a terrible word for what happens. It's more like a *cure*. A *change*… (Goonan 1994, 43)

This putative victim does not want to be saved or fixed or cured. Similarly, Blaze remarks after his Scott Joplin infection that "[i]t's just wonderful, the things that I'm starting to know" (Goonan 1994, 56). What prospectively terrifies is retrospectively wondrous, at least sometimes. When everything is information without evaluative scruples we may be tempted to fall back to consent as the ultimate standard of evaluation, but consent too is compromised in *Queen City Jazz*. The illusion of choice runs deep when our most basic fears and desires are metapheromonally programmable.

Verity's community does become infected, of course, and this begins to foreground the extent to which our individual identities are determined by the form of community in which we live and through which we experience the world. All her life she had been surrounded by people who in a way defined her, told her who she was, and now it was all gone (Goonan 1994, 136). Trying to retain her*self* in the aftermath of the infection, Verity wraps her dead friends in forbidden Enlivenment Sheets and sets out to deliver them to the Queen City, where she hopes they may be revived. She learns, too late, that the Sheets may do something very different from what she intended. Whether her friends will be healed is an open question. They will certainly be changed. Verity agonizes over whether she has made the right choice for them (Goonan 1994, 146). Is any life better than none at all? What can we want for each other (Goonan 1994, 272)? Is it our prospective or retrospective judgment that matters most in contemplating truly radical change (Goonan 290)?

Inside the Queen City Verity discovers a different form of human reality, or perhaps a form that is no longer human (Goonan 1994, 136, 176). Pheromones control the citizens' sense of familiarity, recognition, safety, and well-being. Pheromones turn *no* to *yes* (Goonan 1994, 171). As we readers might have foreseen, their nanoeducation goes right to the heart of volition and sense of self (Goonan 1994, 309). Some citizens of the Queen City are from to time aware that something is wrong, but pheromonal misdirection prevents them from investigating and addressing it.

> [T]here were great ramifications for the memory sponge] in just about every realm of sociological control. They interfaced directly with the brain, and could hold an infinite variety of assemblers and pheromonal analogs…Encyclopedic information flooding into the brain – but *whose* information, and under *whose* control? (Goonan 1994, 312)

They knew in advance that the pheromone breakthrough could turn out to be a cruel disguise for eliciting complete obedience. The alternatives, they thought, were to be controlled by private business concerns or

> [c]onsciousness by committee…A dictatorship of direction. A Knowinger Than Thou conglomeration of social scientists, economists, engineers, and a single, somewhat twisted nanoarchitect [named Durancy]. (Goonan 1994, 319)

Any person on the Committee had the ability to step in and subvert the entire plan, though Durancy was the only person to recognize and exploit that fact (Goonan 1994, 320). Like the original Mother Ann and her cohorts, Durancy recognized the need for a new vision of what human society could be in an Enlivened City, an Enlightened Society. Thanks to nanomedicine they would be free from disease. Thanks to material nanoengineering they would be free from hunger and material deprivation. Thanks to the immense labor savings of the system, no one would need to work. Whatever else it would be, the Conversion would be irreversible. Durancy asked himself what people would *do* in an Enlivened City. *How* would they live?

His ultimate vision was of a society in which "[t]here would at last be time for people to develop creative energies. Their *individuality*…" (1994, 360). He could have left his vision open, as an empowering indeterminate idea of personal freedom, but instead he gave it a determinate form by imposing his own substantive conception of the good. Durancy decided that the Enlivened Queen City should be a city of superlative art, a city in which citizens could thoroughly celebrate art. Through metapheromones, he thought, we would be able to experience art, experience lives of art, as we never had before.

> [It would be] a symbiosis, if it worked. An organic unity with his mind and brain the interface, the consciousness which sensed and would enjoy and savor and live something other than himself, a piece of another's life, more delicious than mere reading, or hearing, or seeing, or touching. Art raised to the nth degree…. He could be *anyone*,

> and then return to himself, like reading a book only immensely more intense. Yes. He could be…*everyone*. (Goonan 1994, 258-9)

In the post-Surge Queen City, citizens perform. Pheromonally immersed in their parts, *living* their parts right down to their DNA, they execute the City's Program (Goonan 1994, 128-9). They live out Flannery O'Connor's *Wise Blood*. Billie Holiday sings nightly. Citizens immortalize Ernest Hemingway, Charlie Parker…and Mark Twain. One might say that the citizens and Bees of the Queen City execute the episodic memory of the City superorganism. Occasionally the original consciousness of a citizen surfaces between parts to briefly to savor the experience, to rejoice in triumph over embodying a coveted role, or to rage against playing an inferior part in the life of the City (Goonan 1994, chapter 10). At least some citizens are able to exercise a level of discretion as to which roles they play, but opting out is not an option. There is no exit from the City. While the Flowers bloom, everyone plays their part.

The principle which makes the ultimate dystopia inevitable in the novel is the natural law that *any form of information transmission suffers loss* (Goonan 1994, 228). At the technical level, the information nanoscientists acquire is perpetually incomplete, thus its products are inevitably incompletely understood and often flawed. These products, the assemblers, are self-replicating. Each replication is subject to additional information loss. The natural cycle from epidemic to endemic to diffuse parasitic diversity and occasional symbiosis is thus replicated in the artificial nanosphere.

At the psychological level, addiction is a byproduct of information loss. Everything is information, including feeling and the intensity of experiences of art. The parts citizens play don't always "take"(Goonan 1994, 208). Even when they do, the margin of return from each artistic performance diminishes for the citizens, and also for the bees who became addicted to the metapheromonal byproducts of human emotion, specifically those of stories, music, and art (Goonan 1994, 228).[13] Chasing that initial high, the Bees, who are perhaps themselves agents, cause the same things to be ceaselessly relived and recycled, each time with additional, vital, loss. Incorporating human limbic tissue both gave the Bees needs that bound them to the City and gave them power to bind humans to their needs.

[13] See Weinstone 1997 for an exploration of the work addiction does in discourses of virtual reality, including in *Queen City Jazz*. Weinstone takes Verity herself to be a predestined addict to the virtual reality of the city. Though she does not directly apply her insights concerning logocentricism and transcendence to Verity, Weinstone's treatment of the complexities of how we think about addiction raises a host of further issues.

One might surmise as well that at least some of the citizens are addicts to performance highs, including the superficial highs of undeserved accolades with dwindling margins of return. Citizens do not write new material, nor do they aspire to originality or genius in interpretation. They are pheromonally determined to be willing vessels, not agents, not artists. For all his utopian vision, Durancy failed to account for attenuation and corruption of the signal. He failed to guarantee room and time for genuinely new, original art. He failed to distinguish between the value of maintaining and preserving old information and the value of originating information. Humans and Bees can both experience art, but only humans can produce it in the Queen City. Without autonomy, the losses of the system are irrecoverable. Consequently Durancy's nostalgic dream of the artistic life was not in the end an opportunity to explore creative energies and develop individuality. It was for many a living nightmare. Nanotechnology freed citizens of the Queen City only to become victims of his unscrupulous philosophy (Goonan 1994, 134).

What I hope to have communicated thus far through this metanarrative is the logic of *Queen City Jazz*, i.e. the principles on which the world and its plot are structured, with an eye to the implications for how we should proceed in designing our future. Every cognitive phenomenon can presumably be exploited for ill or for good. Perhaps there really are clear cases of each, but we know there are also many cases for which we lack robust standards of adjudication because the standards we normally employ presuppose natural kinds, obvious boundaries between health and illness and between the jazzy improvisation of nature and catastrophic metastasis.[14] When the delimitations of our taxonomy are themselves challenged, we may attempt to fall back on informed consent but this too is compromised along with the limbic system, with communication, with information. The kinds of cognitive change we may initiate (even with quasi-magical nan) are typically irreversible, so we cannot test them out, comparing our judgments prior and post, prior and post. When the change is radical, we cannot extend our judgment or imagination to break new ground with any confidence in our accuracy. In *Queen City Jazz*, the power of information can corrupt as easily as it can heal, and we are seldom able to tell the difference.

But what could this novel possibly tell us about the *real* world where we actually live? How could we possibly learn anything from a fictional narrative about a world and its inhabitants so radically different from our own? Appeal to fear taints the experiment

[14] Wolmark identifies *Queen City Jazz* as one of the few science fiction works that challenges us to move beyond the binary (Wolmark 2002, 77).

throughout and it seems there is far too much (irrelevant) detail even in this greatly simplified description for the novel to be a well-defined experiment by any stretch of the imagination. I turn to these issues in the next section before attempting to draw any conclusions from the experiment.

Novels as Thought Experiments

Elgin (2014) and Carroll (2013) have argued that fictional works, including science fiction, can be used as thought experiments. Though there is significant controversy over whether and how anything can be learned from thought experiments, some of their most defensible uses are particularly germane to idea-driven science fiction (Brown and Fehige 2014; Thagard 2014). One use is as a tool for scientific discovery, as opposed to contexts of verification (Schickore 2014; Stuart 2014, 266). In the context of discovery thought experiments can be very effective tools for generating hypotheses and for revealing conceptual shortcomings of theories, e.g. ambiguity of scope, which impact the design of verificational experiments. Another is the use of thought experiments in evaluative contexts, specifically those in which the issue is not how things are but how they ought to be, e.g. in moral philosophy.[15] Thought experiments usefully raise our consciousness of the principles, dispositions and affects we actually employ in moral judgment, e.g. implicit biases, making them available as targets for reflection, critical analysis, and higher order affirmation or rejection (see Cikara *et al* 2010). We sometimes use thought experiments to convey a distinction that we cannot adequately convey by other means. In this section I will focus on how idea-driven science fiction novels like *Queen City Jazz* may legitimately be used as thought experiments for ethics.

Suppose that to argue from a science fiction novel, or from any fictional narrative, is to engage the audience in a thought experiment that effectively "pumps" their intuition:

> Thought experiments are among the favorite tools of philosophers, not surprisingly. Who needs a lab when you can figure out the answer to your question by some ingenious deduction?...Some thought experiments are analyzable as rigorous arguments, often of the form *reductio ad absurdum*… Other thought experiments are less rigorous but often just as effective: little stories designed to provide a heartfelt, table-thumping intuition – "Yes, of course, it has to be so!" – about

[15] Thomson's violinist is perhaps the most famous example in general and the most cited in the discourse concerning the epistemic value of thought experiments (Thompson 1971).

> whatever thesis is being defended. I have called these *intuition pumps*. (Dennett 2013, 6. See also Dennett 1995)

If a given thought experiment does no more than pump one's intuition, i.e. if it does no more than trigger a pumped-up conviction of necessity, then intuition pumps are merely persuasive devices that circumvent evidence based reasoning and pose a rhetorical obstacle to critical reflection and epistemic progress. Such appeals to bare intuition have been widely criticized on a variety of grounds, especially ethical intuition pumps in their radically simplified trolley problem form.[16] The simplest and most general purpose of asking people to choose whether to pull the trolley track lever to save five people thereby killing one, or to allow the trolley to continue on its current track to kill five, is to *prove* that certain distinctions are in fact *universally valid* and *decisive* ethical considerations. It turns out that most people do in fact take distinctions between commissions and omissions, between intended effects and unintended side effects, and other trolley-isolable considerations to be morally relevant. Yet whether such considerations are decisive, whether their weight is individually or culturally relative, and whether we ought to treat such considerations as we in fact do are further questions that a bare appeal to intuition cannot answer. Circularity (a.k.a begging the question or preaching to the choir) and the naturalistic fallacy (attempting to infer an *ought* directly from an *is*) are well known logical hazards of appeals to intuition. Trolley problems usefully raise consciousness of our intuitions, but they do not determine our second order evaluation of the intuitions they reveal.

Trolley problems also face less well known but equally important experimental design challenges. They are specifically designed to isolate one consideration from all others by removing as much context as possible. Though a great many participants in trolley problem thought experiments ask or even demand to know the history of the scenario and details about themselves and those whose fates they are to determine,

[16] For an extended and somewhat biased but popularly accessible account of trolley problems and what they show, see Edmonds 2013. Like many who employ trolley problems in their work, Edmonds assumes that "[t]he point of *any* thought experiment in ethics is to exclude irrelevant considerations that might cloud our judgment in real cases" (Edmonds 2013, xiii emphasis added). I briefly argue below that extended narratives like science fiction novels that make no attempt to control for allegedly irrelevant considerations offer an alternative form of appeal to intuition that is philosophically useful in important ways. For the purposes of this paper I will lump together all highly controlled ethical thought experiments, those which are designed to isolate one consideration by excluding all context that might surreptitiously offer alternative grounds for judgment, thereby contaminating the experiment. Thompson's violin thus counts as trolley problem, whereas *Brave New World* does not.

these sorts of contextual details are withheld. By controlling for context and requiring participants to choose only on the basis of information that they may explicitly judge to be inadequate, requiring them to do so quickly and without benefit of discourse with others, we putatively learn something about the principles that people in their calm considered judgment think they ought to employ. Controlling for context in this way can be criticized as dehumanizing and, drawing from the Kohlberg/Gilligan debate concerning moral development, sexist or androcentric (Blum 1988; Flanagan and Jackson 1987; Schwartzman 2012). Treating the patients of the experiment as mere generic ahistorical and interchangeable bodies on the track may effectively control for partiality but it arguably does so at the expense of dehumanizing the patients (and perhaps thereby hamstringing the experiment by dispensing with an indispensable moral ground). The agents of the experiment are arguably also dehumanized in that the chooser is treated as an ahistorical, rationally ideal, radically ignorant but morally culpable causal power with unnaturally restricted options. Trolley-style problems are also arguably sexist or androcentric (Benhabib 1986. See also Puka 1990). To characterize the gender dispute very crudely, Kohlberg found that there are stages of moral development and that mature men reach the highest (explicitly but nominally Kantian) level of moral development far more often than women (Kohlberg 1973, 631-2; Kohlberg 1981). Gilligan replied that Kohlberg begged the question by presuming that Kantianism is the highest level of moral development (Gilligan 1982).[17] Given that trolley-problems preclude grounds of care and they force participants to choose on Kantian or Consequentialist grounds alone, such experiments may be deeply gender biased in design, thus they may not actually show what they seem to show about moral psychology.

Although it may be true that in extraordinary circumstances humans must actually solve trolley-like problems, these are far from the human norm. Our lives do not consist of a sequence of emergencies that never come to constitute a personal history or a shared

[17] Kohlberg's experiments actually showed that a particular percentage of mature men reason in a particular way and a different percentage of mature women reason in that way. Head counting alone cannot determine whether the minority is defective or ideal or neither. In presuming that the employment of impartial principles is the hallmark of the highest level of moral development Kohlberg presumed that rationalism had won the debate between rationalists like Immanuel Kant and sentimentalists like David Hume. Kohlberg attempts to address this criticism in Kohlberg and Boyd 1973. The movement sparked by Gilligan's insight surprisingly framed the historic debate between rationalists and sentimentalists as posing a false dilemma between agent-centered moral theories. Care ethics is fundamentally a relationship-centered class of moral theories that ultimately challenge the very notions of agents and patients employed by competing moral theories (Held 2005).

history between known participants. Most people do in fact take these mundanities to matter, too. When we generalize from considerations that are really peripheral, we obviously run the risk of seriously distorting or perverting the core of what we wish to investigate. In attempting to generalize from trolley problems and other very simple ethical thought experiments, then, we must rehumanize agents and patients, reinstate their histories and futures, and reconstitute their relationships while avoiding circularity and the naturalistic fallacy. It is not easy to do this well.[18]

The more bizarre thought experiments which have been (mis)taken by some to decisively determine necessary truths are arguably among the most misleading.[19] According to Wilkes, it is precisely because the thought experiment to be executed is inadequately described that the resulting intuitions are ungrounded or go awry:

> [W]hen we have thought experiments in philosophy, there are as we shall see problems in making the inference – precisely because of the ambiguous uncertainty concerning the relevant background conditions, leaving it unclear whether we have 'established a phenomenon'. This means that our intuitions run awry, and the inferences are not only problematic, but the 'jump' from the phenomenon to the conclusion is made the larger because of the further need to imagine just what these backing conditions, under the imagined circumstances, would be. The 'possible world' is inadequately described. (Wilkes 1994, 8)

One possible advantage of appealing to intuition using a science fiction novel is that the extended narrative of a novel deeply contextualizes all the participants in a shared history in a counterfactual world. When a reader becomes episodically immersed in the imagined circumstances to the extent that she is affectively engaged and invested in what happens, perhaps even mirroring the characters, we have some grounds for claiming that the possible world is adequately described.

Instead of controlling for context to isolate a single declaratively expressible consideration, then, the text of a novel engages our episodic faculties. Episodic

[18] Gendler attempts to explain how peripheral or exceptional cases can ground generalizations (or universalizations) by distinguishing between norm-driven exceptions and exception-driven norms. She argues that thought experiments (always) operate by appealing to exceptional cases that drive norms by revealing particular regularities to be non-accidental, deeper truths about the world (Gendler 2000, xii, 142-3, 150ff).

[19] Thagard for example argues against ever using intuition pump thought experiments, especially in cognitive science (Thagard 2014).

memory, episodic foresight (sometimes called mental time travel), and episodic mindreading (sometimes called theory of mind) are experience-building, world-building, *simulational* activities that can engage us all the way to the visceral level. Immersion in a narrative is an episodic activity that recruits the same brain structures to simulate counterfactual experience. Each of these can be used for entertainment, but they also play an indispensable role in the *systemic* construction of our Ends. At an individual level, episodic memory and foresight help us learn from our successes and failures, develop our characters, and construct life plans, in part by helping us work out how multiple relevant considerations do and should interact in a complex world. At a social level, episodic mindreading or empathy is critical for the development of relationships, the construction and pursuit of shared ends, and the authentic inclusion of real others in one's lived experience.

The obvious disadvantage to the extended narrative thought experiment is the lack of control. The individuals who participate in the experiment may differentially attend to details, so they may effectively be judging on different bases. Given that even the most detailed narratives underdetermine their full simulation in imagination, the details each participant creates to fill the world may differ in experimentally relevant ways. From characters and plots of any significant complexity, many different kinds of conclusions may be drawn. It might well turn out that most narratives are simply inconclusive, supposing that the purpose of the experiment must be to determine universally valid and declaratively communicable results. Perhaps most devastatingly for Western analytic philosophers, this kind of argument is not reducible to a standard argument form that would be logically compelling from the third person perspective. Episodic reasoning is fundamentally first person, only partially expressible in declarative form, and only some of the inferences are logical in nature. Whereas trolley problems must avoid circularity, naturalistic fallacy, dehumanization, and androcentricism, extended narratives must avoid subjectivity, inconclusiveness, incommunicability, and incommensurability.

These pitfalls of extended narrative thought experiments are mitigated by several factors. Unlike stereotypical controlled experiments, simulations are not designed to be conclusive upon initial performance. Simulations are designed to be run repeatedly, so that the effects of chance can be modeled and the parameters can be varied to explore the dynamics of the system. Episodic simulations like rehearsing a gymnastic routine or an important upcoming social interaction are likewise most effective when repeated over time with a field of variations. Their purpose is seldom merely to predict what will happen. Episodic simulations are empowerment mechanisms that serve to increase our influence in events by helping us understand and plan for a range of contingencies in which we participate. Since many of the contingencies for which we need to plan are determined by the choices of others, episodic simulations are often most effective when performed and critically evaluated in cooperation with others, e.g. mock interviews. Insofar as an idea driven science fiction novel is an episodic simulation for the reader, then, we may expect to learn more from the thought experiment upon repetition, perhaps over a period of years, and in company, e.g. in a class, a book club, or an internet based fan community. When a narrative is so widely shared and repeated that it becomes part of our cultural heritage (e.g. *Star Trek*), the pitfalls of subjectivity, incommunicability, incommensurability, and inclusiveness may become negligible.

Applying these considerations to *Queen City Jazz*, the most obvious hazard of treating this novel as a thought experiment is that our intuitions are clearly being pumped in ways that may compromise the experiment. There are several morally relevant intuitions that the author overtly uses to generate tension and suspense to drive the plot, many of them fears. As a component of a general fear of science run amok, Goonan plays on our fear of unseen (nanoscale) dangers and our fear of the unknown, unanticipable, irreversible consequences that science makes possible. These fears are very common drivers of science fiction drama. Goonan also plays on fears that are typically more specific to first contact and plague-apocalypse science fiction, namely our fear of change, particularly of being changed, as the exogenous becomes endogenous. In this vein she plays on our fear of violations of our personal boundaries from our skin to our will. All these fears shape the evaluative field of judgment both within the novel and to a lesser extent in the real world. By inculcating these fears in the reader, Goonan shapes the ends the reader attributes to the characters as well as the ends the reader wants for the characters. With repetition and reinforcement these fictionally contextualized fears eventually leach into how readers experience the real world, thus how we shape our real ends. Memory, dreams, fantasies, and current experience are less dichotomous and less discrete than many of us assume. Fiction can fundamentally change us by inducing vicarious trauma or inspiration. The affects (feelings, emotions, etc.) may attenuate rapidly but the lessons we learn may be quite lasting (cf. *The Grapes of Wrath, Brave New World, The Help*, etc.).

If *Queen City Jazz* is to be anything more than an intuition pump, it must adequately describe a coherent cognitive process that *shows* something intersubjectively valid. In particular, the characters must be adequately described *subjects* in whose shoes we can walk an imaginative mile. Tampering with the limbic system is of course an obvious threat to subjectivity but no more so than the cybernetic implants or telepathic control described in other science fiction novels. What makes *Queen City Jazz* distinctive, thus more worrisome, in its approach to subjectivity is that Goonan explicitly recognizes the power of fiction and other arts to fundamentally change who we are, and extends this idea to its extreme in the Queen City. The actors *become the part* as literally as we can imagine. The distance between act and audience is blurred or eliminated, as they *become* the extras of the scene, embedded in the times and places they relive. An inadequate description of the citizens could easily make them incoherent as persons, into non-subjects we can consider only as objects.

The distinction between the lived part and the life is thus crucial to the adequacy of description in Goonan's appeal to intuition. Without it, citizens of the Queen City might seem too alien for our empathy to engage. Appeals to intuition regarding the nature and value of subjectivity would then fall flat. Goonan solves this problem by protecting the continuity of the real subject as a substrate for the lived parts, a substrate that surfaces between parts. Some citizens play a variety of short parts with time off between them. These citizens are only one step removed from contemporary method actors. Other citizens might be immersed in a single part for great lengths of time, but Goonan provides an off season to make the lived part discontinuous. During the off season when the Flowers become dormant in the winter, the citizens have an opportunity to recover or reconstruct themselves. Without this time-out from living the artistic life Durancy envisioned, it is not at all clear that there would be any substrate of an individual agent left to recover or any rational will left to do the reconstructing.[20] The subjectivity of the characters is thus deeply compromised, but not annihilated by complete immersion in the *Queen City* life. The off-season serves to preserve the reader's intuition that these are people, still humans thus moral patients if not agents, living a largely inhuman (deeply wronged) life. We can simulate what it would be like to be one of them, feel the problem,

[20] Wolmark makes a similar point in terms of a normative or natural "unitary" human subject that may be conjoined with or interpenetrated by a technological other (Wolmark 2002, 77). Following Hayles, Wolmark takes the defining environment for the contemporary technologized body to be that of the separability of form and matter and the identification of the human as formal rather than material (Wolmark 2002, 78). Though this separability is prevalent in *Queen City Jazz* and Wolmark is correct that in much of the genre this "entails a loss of social, cultural, and sexual specificity," it is noteworthy that mundane social relations, local culture, and the body are left almost entirely intact in *Queen City Jazz*. Citizens are put to work involuntarily, their bodies and minds are *used* on a regular basis, but they have lives outside the job and their bodies are left almost untouched. A few characters have nubs or a glow. One has paws. One becomes a Bee. Though they presumably could make an art of body modification or more radical self-change, the characters do not even fiddle with their skin color or secondary sexual characteristics, much less make themselves beautiful or monstrous. It is almost exclusively the subjectivity of the characters that is at risk, i.e., vulnerable to exogenous subjectivities.

and reflect upon it. It remains to be seen what we might learn from doing so. To this we turn in the final section.

What Follows from *Queen City Jazz*

Supposing for the sake of argument that its experimental design can be validated despite the problems raised above, what may we learn from *Queen City Jazz*? If Goonan's appeal to intuition works, we readers immersed in the narrative should presumably become at least temporarily very concerned about scientists who attempt to tamper with our limbic system, or even with more moderate attempts to enhance and deepen our virtual reality experiences without biological tampering. More mundanely, we should be very concerned with how our immersion in commercial media may be tampering with who and what we are. The lasting message from fear is that we should be very careful what we wish for. These are, of course, merely intuition pump results. If the experiment really is well designed, it should be possible to diagnose the specific problems in the novel, and by articulating what has gone wrong, really learn something about the values that structure coherent ideals for our future. In this section I describe the protagonist's diagnosis and solution, then interpret the lesson in Marxian terms before arguing that the novel shows something that is very difficult to tell, that humans need, thus ought, to work.

What specifically makes the Queen City or its world dystopic? The real problem cannot be that things change or that people change. Growth and development are *ex hypothesi* good even though they are irreversible, often surprising, and unavoidable. The problem cannot be that some of the efficient causes are too small to see, nor can it be that they are exogenous. The human super-organism, complete with complex microbiome, is a constant flux of microscopic interaction and environmental exchange. We do not live in fear of these.

The general diagnosis Goonan explicitly offers is, not surprisingly, that citizens of the Queen City are not free. There is little volition in the city (Goonan 1994, 352). Verity eventually works out that she is the new Queen and her choices will determine the future for everyone. It is her problem to determine how to free the city from its cycle, and the reader in her protagonist shoes is expected to hypothesize a variety of possible solutions before Verity's solution is revealed. According to Verity's diagnosis, the problem is that the city is a closed system that has been overengineered to serve a unitary purpose with no need for human maintenance or room for growth and development. Her first act is to give citizens a choice to opt out of the city life. She knows that merely opening the doors will not really enable anyone to escape their addiction – too few will *want* to leave – so she infects the population with a plague virus designed to force them out of the city and then wear off. This both opens the system and gives citizens a new perspective on their options. Her second act is to reseed the city with a "less is more" pluralistic ideal (Goonan 1994, 460). Verity's vision is of a sane and functional engineer's city *with* art but not *for* art (Goonan 1994, 450). We are given to believe that her solution is adequate because even India, the monstrous mother[21] who is putatively the heart of the problem, finds it freeing:

> She watched an amazing change come over India's face. Terror, sorrow, grief, anguish, and then joy suffused her features in quick succession, and then a puzzled wonder as a smile appeared and tears began to flow. Sobbing, she approached Verity, and Verity could not move.
>
> India embraced her.
>
> "Thank you," she whispered. Her face was growing old, into the face Verity had found so dear. "I thought I never could be free." (Goonan 1994, 453)

Verity's plan is to make room for growth and development by *freeing* the citizens both *from* the psychological and physical constraints that imprison them within the city and *from* the commodification of their artistic labor.

The underlying Marxian point here, that a city *for* art, especially one run by Bees addicted to its products, would be a city designed to alienate its citizens from their creative powers, should not be lost on anyone. Like Mother Ann, Ma, and Durancy, Marx and Engels were social architects. Like Goonan and her character Durancy, Engels saw that new technology has the potential to either produce misery and crisis, as big industry did in his day, or "in a different form of society" to free us:

> …large-scale industry and the unlimited expansion of production which it makes possible can bring into being a social order in which so much of all the necessities of life will be produced that every member of society will thereby be enabled to develop and exercise all his powers and abilities in complete freedom. (Engels 1847b, 347)

[21] See Kornfeld 2004 for more on the monstrous mother in *Queen City Jazz.*

Unlike the nanoarchitects of *Queen City Jazz*, however, Marx had no intention of attempting to free humans from working for their own sustenance. He conceived of labour, human power, and creativity almost exclusively in terms of production for natural sustenance. His aim was to organize society such that the means of securing sustenance are held in common rather than privately held, so that no one would be excluded from or have to compete for a fair share of the benefits of mass production. Unlike Durancy, Marx did not in general deem it a bad thing for a human to live by his own labour. His issue was with labouring for others under threat of material insufficiency, as serfs, slaves, and proletarians must:

> We by no means intend to abolish this personal appropriation of the products of labour, an appropriation that is made for the maintenance and reproduction of human life, and that leaves no surplus wherewith to command the labour of others. All that we want to do away with is the miserable character of this appropriation, under which the labourer lives merely to increase capital, and is allowed to live only in so far as the interest of the ruling class requires it. (Marx and Engels, 1848, §II ¶39, 499)

The fundamental problem Marx aimed to solve was the commodification of labour, i.e. the reduction of the value of human activity to an exchange value. Thinking along these lines, one might diagnose the fatal flaw of the Queen City as an incomplete decommodification of labour. Whereas all other commodities became free, i.e. they required no exchange or return for their use, the pheromonal byproducts of our experiences as citizens became a new commodity. Worse, our enslavement to the production of this new commodity was such that our compliance with the social demand for production was involuntary in entirely new way. Our bodies and minds are fundamentally *used* as mere means in the Kantian sense[22] to satisfy the needs of the system on which they depend for sustenance and from which they cannot escape. If Verity's plan succeeds in bringing genuine creativity back to the city, the Bees might no longer suffer from information loss. If engineers populate the city, perhaps even the Bees could be freed, but freed to do what?

[22] Kant defined prudence as skill in using others as means and argued that prudence is an important step on the path towards moralization, but he most famously argued that using others as "mere" means nevertheless violates the formula of humanity (Kant 1803, 9:450; Kant 1785, 4:429).

Despite their deep occupation with the productive, Marx and Engels occasionally described freedom in terms unconstrained by productive purpose, intimating perhaps that they too were ultimately concerned with freedom *to*. In *A Communist Confession of Faith*, for example, Engels articulated the central aim of the Communist party in terms of the sustainable free exercise of human powers with no reference to production:

> [The aim is to] organise society in such a way that every member of it can develop and use all his capabilities and powers in complete freedom and without thereby infringing the basic conditions of this society. (Engels 1847, 96)

Perhaps Marx offered no positive account of complete freedom because he, like Kant, believed that we can only discover what we may become through the progress of history.[23] Engels did at least envision that the systemic change he advocated, communism, would require us to become "quite different people" (Engels 1847b, 353). Communal control over production would put an end to hyperspecialization, he argued, because the kind of planning it requires "presupposes moreover people of all-round development, capable of surveying the entire system of production…[free] from that one-sidedness which the present division of labour stamps on each one of them" (Engels 1847b, 353). Whether or not Marx and Engels were correct that communism would have the general effect of making us all better-rounded, we should take the point that the reciprocal influence between individual development and social development must figure prominently in our planning for the future. It would be a mistake to take how people are as a given that drives what society may be for humans. We humans rise and fall to the occasion, depending upon what is demanded of us and what resources we may bring to meet those demands.

When the demands are lifted and resources remain, what we would *do* is an open question. What we would *be* as subjects or agents or persons is likewise an open question. In an individualistic capitalist society, it would not be surprising to find that many or most readers of *Queen City Jazz* diagnose the fundamental problem of subjectivity in Goonan's world in terms of losses and gains of ownership, or in more Marxian terms, estrangement from what should be one's own.[24] On this interpretation the problem is that what ought to be mine is not really my own, or that my ownership of what ought to

[23] See Dupré 1998 for a fair interpretation of Kant's theory.

[24] Marx is perhaps as famous for his 1844 essay "Estranged Labour" as for co-authoring the Communist Manifesto (Marx 1844, 270).

be deeply mine is somehow compromised. When an exogenous metapheromone package changes my *no* to *yes*, what matters not that it came from somewhere outside me, but that I act on it without owning it. I am estranged from my own labor, as Marx would say, at the deepest psychological level. I may accept the inevitable as a slave complies under coercion, but the *yes* remains alien. The exogenous cause thus remains a trespasser in my will. My right to reject the exogenous from the domain of what is most essentially mine, by body, my thoughts, my will, is deeply compromised on this view.[25] This ownership model of what is *wrong* with limbic tampering is, as I have described it, a rights based understanding of the intuition generated by the narrative: My claim to the use of my mind and body is impotent.

Ownership in this broad sense is clearly a morally significant feature of cognition, and it is one with very deep cognitive roots (Shaw et al 2012; Kalckert and Ehrsson 2012; Limanowski 2014). To give a few cursory examples, Mirrors can allow an amputee to scratch the itch in an absent limb, arguably by creating an illusion of ownership that satisfies the relevant body mapping demands. Schizophrenia is in part characterized by thoughts and desires that are experienced as exogenous and alien, not one's own (Martin and Pacherie 2013). The "first-personness" of episodic memory and episodic foresight can be compromised such that one can remember episodically only as if it happened to someone else, or imagine *some*one's future though not one's own, and this significantly compromises agency (Martin-Ordas *et al* 2012). Ranging more widely, the endowment effect in behavioral economics (also known as loss aversion or divesture aversion) reflects how perception of ownership influences judgments and behaviors in neurotypical agents (Shu and Peck 2011).

Even granting that the ownership component of subjectivity is a universally valid morally relevant consideration, I contend that the *Queen city Jazz* experience machine *shows* that *work* is also a fundamental component of subjectivity and a prior one at that. Like ownership and the mineness of my body, thoughts, and choices, *exercise* or *work* has deep biological and cognitive roots. From the adage that "neurons that fire together wire together" to cardiovascular exercise and pedagogy, the indispensability of mental and physical purposive exercise to the healthy development of a human being is widely recognized. We see it in toddlers who begin to reject aid in order to "do it myself," however ineptly. We see it in the charges of infantilization and disrespect laid at

[25] Mark Huston's talk "Black Mirror's 'The Entire History of You': Memory as a Recording Device" at *The Work of Cognition and Neuroethics in Science Fiction* conference held by the Center for Cognition and Neuroethics in March 2015 helped sharpen my thoughts on ownership of one's memories.

the doors of helicopter parents, overbearing partners and paternalistic politicians who presume to make things too easy for us. Even when our ends agree, it is often critical to me that *I paint* the painting, *I earn* the income, *I play* with my child, or *I choose* the gift. What is important is not *that* it get done or *that* it be done well. The do*ing* is constitutive of the end. A prioritization of work is clearly recognizable in today's Maker movement and in the lives of athletes and musicians who take the exercise of their skills to be non-instrumental ends that structure a way of life.

We have long seen the combined indispensability and priority of work in philosophical arguments valuing activity over passivity or advocating agent-centered moralities over patient-centered ones. From Aristotle to Kant, Marx, and contemporary race theory, philosophers and critical thinkers have long recognized the indispensability of work to human nature, human development, virtue, and happiness. Simplifying their views to an extreme, Aristotle argued that virtue is rational *activity* and the life of rational activity is the life of *eudaimon*. Kant argued that autonomy, the cognitive exercise of freedom, is the good upon which the value of all other goods depends. Marx argued that humans are fundamentally *laborers* who generate value through the exercise of our human capacities. Recently feminists, critical race theorists, and even business ethicists have argued that meaningful work is a fundamental human need that generates moral protections against legal and social exclusion from work, as well as rights to maximal autonomy and freedom of expression in workplaces.[26]

Some of these indispensability claims can be reframed in terms of ownership – my *agency* is after all *my* agency – but we should take care to avoid recklessly reducing what I *do* or what I *am* to what I *have*. The verbs differ. Advocates of the priority of ownership, including those who grant that exercise is indispensable, may contend that self-ownership is metaphysically prior to exercise. There has to be a me in order for me to be an agent, one might argue, and mineness is constitutive of me. This can be resisted by distinguishing between metaphysics and metacognition. Though most Western analytic philosophers may take it as settled that *being* is metaphysically prior to *doing*, this is not the only coherent metaphysical position. More importantly, even if we grant the priority of being over doing, it does not follow that ownership is prior to exercise. Ownership or mineness in the relevant sense is a psychological or cognitive category that may not be reducible to a metaphysical or epistemic category. What matters in the first person is

[26] Iris Young for example has argued forcefully that exclusion from work, i.e. marginalization, is a largely unrecognized and highly dangerous form of oppression (Young 1998).

not necessarily whether the thought *is* mine or even whether I *know* it is mine.[27] What matters in the first person may instead be whether I *own* it, whether I *take* it to be mine, or whether I *identify* it as alien or *reject* it. Mineness in the sense relevant to *Queen City Jazz* is thus itself a cognitive activity, a stance one takes towards a body, a thought, or an activity. Insofar as ownership can be a *stance one takes* (or a stance that one is unable to take) towards one's own thoughts and activities, ownership is itself a metacognitive activity[28] rather than a metaphysical relation.

Returning to *Queen City Jazz*, the novel clearly supports the view that work is diagnostically critical. Throughout the novel there are hints that development and growth require work. There are constraints on the ways in which certain important kinds of information can be acquired. Maturity, mastery, and creativity cannot simply be given. For example, Verity must be annually programmed to guide her growth. Children in the Queen City learn metapheromonal programming the hard way before they are allowed to let the city do it for them. Human development always requires something endogenous:

> …They [parents] try and tell you things, important things that they've learned…But soon they learn that they can't, not really. They can only give you information that is, in a way, oblique. Parents – *good* parents – realize that there are certain things that you have to learn for yourself. It's the act of incorporation that's important. That's what lays down the synaptic paths, not just *hearing* about something. It's *your* doing, *your* failing, *your* actions, your own enormously individual kinesthesis within the world, within matter's confines and matter's release…that causes growth. (Goonan 1994, 377)

[27] I may understand perfectly well that I have schizophrenia and that I experiences some of my own thoughts as alien, but this really doesn't solve my problem. Knowing that a seemingly alien thought is really my own does nothing to reduce its alienness or grant me control of it. Knowing that it's mine doesn't make it mine in the relevant sense.

[28] The distinction between cognition and metacognition is often characterized by the differences between 1) knowing something in the ordinary case, 2) knowing that you know something without being able to recall it (tip-of-the-tongue phenomenon), and 3) recalling something that you didn't know you knew. Ordinarily cognition and metacognition occur as a package deal, as in 1. Sometimes metacognition occurs without cognition, as in 2 where metamemory is disassociated from memory. In 3 cognition occurs without metacognition, as in cases of blindsight or more commonly as in cases of retrograde amnesia in which one discovers that one plays piano or speaks a second language fluently. The conscious reflective first person perspective is metacognitive.

The novel's exemplar of the fully grown citizen, Sphere, is a true musician. He soaks up all the greatness of the past, masters the information, and creates genuinely new works of art.

> …This is a place where you can truly learn things, if you're a part of it…Charlie Parker had this great breakthrough, you know. He was thrown off the stage one time when he was just a kid and then he was absolutely determined to show them. So he went home and played all the records that he could find, over and over, like a maniac. He learned all that and then he tossed it aside. He broke through. He *created*. I think that here you can do that quickly. Learn all the masters that way, then break out into *yourself*, your true self, and still use all that stuff. (Goonan 1994, 386-7)

Passages like this tell us that the exercise of human powers, the kinesis, is indispensable to the development of a human life but they do not clearly disentangle ownership from work. My *doing* it and *my* doing it are merely shifts of emphasis.

The perhaps untellable lesson that the novel shows the reader is that we humans must work to become our selves. The novel as experience machine shows the reader how *we are* and *we become* what we *do*. By effectively stipulating that citizens cannot feel alienated from their activity, cannot reflect upon the activity in out-of-character ways while immersed in it, and cannot even entertain the thought that they are engaged in a performance, Goonan precludes the experienced alienness *in the moment* that is requisite for failures of ownership to be the most fundamental problem. A reader who does the experiment and imaginatively walks in the shoes of a Queen City citizen should consequently find herself demanding the freedom to try it herself, however ineptly, so that she can learn how and make it her own. We first become mature subjects who are capable of owning and disowning through the exercise of our powers.

Science fiction can be useful to the advance of technology not only by helping us envision what is possible but even more importantly by helping us mindfully determine the ends of technology for humans. The daily practice of Western research is not visionary in nature. Scientists and developers are caught up in particular experiments with highly localized ends that primarily concern the incremental extension of knowledge and development of means with little consideration of the ultimate ends to which these may be employed. In order to consider how the advancement of science might best serve our mental health, individually and collectively, we must first determine what we ought to count as health and what role such considerations as ownership and work ought to play in our construction of mental health as an End. We need scruples to do so. The extent to which science may eventually allow us to realize virtual reality, whether via computers or metapheromones, is still an open question. Whether we ought to do something that we can do is a very different question. Considering how radically an individual might be changed, how radically human civilization might be changed, if we plug along with our noses to the grindstone without looking up to see what we will gain and lose, we have good reason to generate and use idea driven science fiction in the construction of our Ends. What we will *do* must be a central consideration.

To make the implications of the pro-work argument more concrete and immediate, consider that the mere fact that it is difficult for autistic people to determine what others are thinking and feeling does not determine whether we (doctors, parents, etc.) ought to try to make it easy. Even if we someday could make it easy for anyone, there might be greater value in figuring it out oneself, however imperfectly. It might do us all some good, individually and socially, to work harder at communicating clearly, accurately, and selectively. Whether the social aspect of autism ought to be a target for medical intervention, now or ever, depends upon a great many interdependent considerations that we perhaps ought to explore through shared narratives, science fictional and otherwise (e.g. Moon 2004; Gerland 2003).[29] We may simply not yet know what is really relevant. Taking the work out of human life might be an enormous mistake. If anyone still has doubts, I suggest you read some science fiction.

[29] The idea is that fictional works like *The Speed of the Dark* might reshape and refine the goals of autism research (Moon 2004). The insistence that autism must be cured is driven in large part by the inability of neurotypicals to imagine what it's like to be autistic, which greatly hinders their ability to find value in an autistic way of life. Rather than adopting the simplistic aim to cure autism we might instead aim to counter only its commonly attendant intellectual disabilities that impair self-help, leaving the core autistic self to her own devices. Alternately, we might aim to cure neurotypicals of their burdensome social needs, their honesty disability, or their abusive tendencies. Before we get ahead of ourselves, of course, we should think very hard about the people we aim to make ourselves and the world we aim to create.

References

Bechtel, W. 1988. *Philosophy of science: An overview for cognitive science*. New Jersey: Lawrence Erlbaum Associates, Inc.

Benhabib, S. 1986. “The Generalized and the Concrete Other: The Kohlberg-Gilligan Controversy and Feminist Theory.” *Praxis International* 5: 402–24.

Boyd, R. D. and J. G. Myers. 1988. “Transformative education.” *International Journal of Lifelong Education* 7 (4): 261–84.

Brown, J. R. and Y Fehige. 2014. “Thought Experiments.” *Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), edited by E. N. Zalta. http://plato.stanford.edu/entries/thought-experiment/.

Buchanan, H., L. Marksonm, E. Bertrand, S. Greaves, R. Parmar, and K. B. Paterson. 2014. “Effects of social gaze on visual-spatial imagination.” *Frontiers in Psychology* 5: 1–7.

Camilleri, K. 2014. “Toward a Constructivist Epistemology of Thought Experiments in Science.” *Synthese* 191: 1697–1716.

Carroll, N. 2013. “Science Fiction, Philosophy and Politics: ‘Planet of the Apes’ as a Thought Experiment.” *Ethical Perspectives* 20 (3): 477–93.

Carter, C. S. 2014. “Oxytocin Pathways and the Evolution of Human Behavior.” *Annual Review of Psychology* 65: 17–39.

Catani, M., F. Dell’Acqua, and M. T. De Schotten. 2013. “A Revised Limbic System Model for Memory, Emotion and Behaviour.” *Neuroscience & Biobehavioral Reviews* 37 (8): 1724–7.

Cikara, M., R. A. Farnsworth, L. T. Harris, and S. T. Fiske. 2010. “On the wrong side of the trolley track: neural correlates of relative social valuation.” *Social Cognitive & Affective Neuroscience* 5 (4): 404–13.

Coskun, A., M. Banaszak, R. D. Astumian, J. F. Stoddart, and B. A. Grzybowski. 2012. “Great expectations: can artificial molecular machines deliver on their promise?” *Chemical Society Reviews* 41 (1): 19–30.

De Bo, G., S. Kuschel, D. A. Leigh, B. Lewandowski, M. Papmeyer, and J. W. Ward. 2014. “Efficient assembly of threaded molecular machines for sequence-specific synthesis.” *Journal of the American Chemical Society* 136 (15): 5811–4.

Dennett, D. C. 2013. *Intuition Pumps and Other Tools for Thinking*. New York: W. W. Norton & Company.

———. 1995. "Intuition Pumps" In *Third Culture: Beyond the Scientific Revolution*, edited by J Brockman. New York: Simon & Schuster.

Dodd, J. and S. Stern-Gillet. 1995. "The Is/Ought Gap, The Fact/Value Distinction and the Naturalistic Fallacy." *Dialogue* 34 (04): 727–46.

Dölen, G., and R. C. Malenka. 2014. "The Emerging Role of Nucleus Accumbens Oxytocin in Social Cognition." *Biological Psychiatry* 76 (5): 354–5.

Doty, R. L. 2014. "Human Pheromones: Do they exist?" In *Neurobiology of Chemical Communication*, edited by C. Mucignat-Caretta and R. L. Doty. Boca Raton (FL): CRC Press.

Doyle, S. M., O. Genest and S. Wickner. 2013. "Protein Rescue from Aggregates by Powerful Molecular Chaperone Machines." *Nature Reviews Molecular Cell Biology* 14 (10): 617–29.

Du, G., E. Moulin, N. Jouault, E. Buhler, and N. Giuseppone. 2012. "Muscle-like Supramolecular Polymers: Integrated Motion from Thousands of Molecular Machines." *Angewandte Chemie International Edition* 51 (50): 12504–8.

Dupré, L. 1998. "Kant's Theory of History and Progress." *Review of Metaphysics* 51 (4): 813–828.

Edmonds, D. 2013. *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton: Princeton University Press.

Engels, F. (1847) 1976. "Draft of a Communist Confession of Faith." In *Marx Engels Collected Works, Vol. 6*, translated by J. Cohen. London: Lawrence & Wishart Ltd.

———. (1847b) 1976. "Principles of Communism." In *Marx Engels Collected Works, Vol. 6*. London: Lawrence & Wishart Ltd.

Fenga, T., W. Zhao, and G. F. Donnay. 2013. "The Endowment Effect Can Extend from Self to Mother: Evidence from an fMRI Study." *Behavioral and Brain Research* 248 (1): 74–9.

Flanagan, O., and K. Jackson. 1987. "Justice, Care, and Gender: The Kohlberg-Gilligan Debate Revisited." *Ethics* 97 (3): 622–37.

Frank, J. (Ed.). 2011. *Molecular Machines in Biology: Workshop of the Cell*. Cambridge: Cambridge University Press.

Gallagher, S. 2013. "The Socially Extended Mind." *Cognitive Systems Research* 25: 4–12.

Gandevia, S. C., L. R. Wilson, J. T. Inglis, and D. Burke. 1997. "Mental Rehearsal of Motor Tasks Recruits α-motoneurones but Fails to Recruit Human Fusimotor Neurones Selectively." *The Journal of Physiology* 505 (1): 259–66.

Gerland, G. 2003. *A Real Person: Life on the Outside*. London: Souvenir Press.

Gendler, T. 2000. *Thought Experiment: On the Powers and Limits of Imaginary Cases*. New York: Garland Publishing.

Gilligan, C. 1982. *In a Different Voice*. Cambridge: Harvard University Press.

Haldeman, J. 1974. *The Forever War*. New York: St. Martin's Press.

Held, V. 2005. *The Ethics of Care: Personal, Political, and Global*. Cambridge: Oxford University Press.

Kalckert, A. and H. H. Ehrsson. 2012. "Moving a Rubber Hand that Feels Like Your Own: A Dissociation of Ownership and Agency." *Frontiers in Human Neuroscience* 6: 40–60.

Kant, I. (1803) 2007. "Lectures on Pedagogy." In *Anthropology, History, and Education*, translated by R. B. Louden. New York: Cambridge University Press.

———. (1785) 1998. *Groundwork of the Metaphysics of Morals*. Translated by C.M. Korsgaard. New York: Cambridge University Press, 1998.

Kohlberg, L. 1973. "The Claim to Moral Adequacy of Highest Stage of Moral Judgment." *Journal of Philosophy* 70: 630-46.

———. 1981. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice.* Essays on Moral Development, volume 1. New York: Harper and Row.

Kohlberg, L. and D. Boyd. 1973. "The Is-Ought Problem: A Developmental Perspective." *Zygon: Journal of Religion and Science* 8: 358–71.

Kornfeld, S. 2004. "Suppression and Transformation of the Maternal in Contemporary Women's Science Fiction." *Extrapolation* 45(1): 65-75.

Lattal, K. M., and M. A. Wood. 2013. "Epigenetics and Persistent Memory: Implications for Reconsolidation and Silent Extinction Beyond the Zero." *Nature Neuroscience* 16 (2): 124–9.

Limanowski, J. 2014. "What can body ownership illusions tell us about minimal phenomenal selfhood?" *Frontiers in Human Neuroscience* 8: 946–52.

Longino, H. E. 1987. "Can There be a Feminist Science?" *Hypatia* 2 (3): 51–64.

Martin, J. R. and E. Pacherie. 2013. "Out of Nowhere: Thought Insertion, Ownership and Context-Integration." *Consciousness and Cognition* 22 (1): 111–22.

Martin-Ordas, G., C. M. Atancea, and A. Louwa. 2012. "The Role of Episodic and Semantic Memory in Episodic Foresight." *Learning and Motivation* 43: 209–19.

Marx, K., and F. Engels. (1848) 1976. "Manifesto of the Communist Party." In *Marx Engels Collected Works, Vol. 6*. London: Lawrence & Wishart Ltd.

Marx, K. (1844) 1975. "Estranged Labor." In *Marx Engels Collected Works, Vol. 3*. London: Lawrence & Wishart Ltd.

Molfese, D. L. 2011. "Advancing Neuroscience through Epigenetics: Molecular Mechanisms of Learning and Memory." *Developmental neuropsychology* 36 (7): 810–27.

Moon, E. 2004. *The Speed of the Dark.* New York: Ballantine Books.

Puka, B. "The Liberation of Caring: A Different Voice for Gilligan's 'Different Voice'." *Hypatia* 5 (1): 58–82.

Sanbonmatsu, K. Y. 2012. "Computational Studies of Molecular Machines: The Ribosome." *Current Opinion in Structural Biology* 22 (2): 168–74.

Sandamirskaya, Y., S. K. Zibner, S. Schneegans, and G. Schöner. 2013. "Using Dynamic Field Theory to Extend the Embodiment Stance toward Higher Cognition." *New Ideas in Psychology* 31 (3): 322–39.

Schickore, J. 2014. "Scientific Discovery" *Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), edited by E. N. Zalta. http://plato.stanford.edu/entries/scientific-discovery/.

Schwartzman, L. H. 2012. "Intuition, Thought Experiments, and Philosophical Method: Feminism and Experimental Philosophy." *Journal of Social Philosophy* 43 (3): 307–16.

Shaw, A., V. Li, and K. R. Olson. 2012. "Children Apply Principles of Physical Ownership to Ideas" *Cognitive Science* 36 (8): 1383–1403.

Shu, S. B. and J. Peck. 2011. "Psychological Ownership and Affective Reaction: Emotional Attachment Process Variables and the Endowment Effect." *Journal of Consumer Psychology* 21 (4): 439–52.

Slotnick, S. D., W. L. Thompson, and S. M. Kosslyn. 2012. "Visual Memory and Visual Mental Imagery Recruit Common Control and Sensory Regions of the Brain." *Cognitive Neuroscience* 3 (1): 14–20.

Stanley, M. 2014. "From Ought to Is: Physics and the Naturalistic Fallacy." *Isis* 105 (3): 588-95.

Steinbeck, J. 1939. *The Grapes of Wrath*. New York: Viking Press.

Stuart, M. T. 2014. "Cognitive Science and Thought Experiments: A Refutation of Paul Thagard's Skepticism." *Perspectives in Science* 22 (2): 264–87.

Südhof, T. C. 2013. "A Molecular Machine for Neurotransmitter Release: Synaptotagmin and Beyond." *Nature medicine* 19 (10): 1227–31.

Taylor, E. W., and Cranton. (Eds.). 2012. *The Handbook of Rransformative Learning: Theory, Research, and Practice*. San Francisco: Jossey-Bass.

Thagard, 2014. "Thought Experiments Considered Harmful." *Perspectives in Science* 22 (2): 288–305.

Thomson, J. J. 1971. "A Defense of Abortion." *Philosophy and Public Affairs* 1: 47–66.

VanHook, A. M. 2015. "Epigenetics of Addiction." *Science Signaling* 8 (366): ec50.

Weinstone, A. 1997. "Welcome to the Pharmacy: Addiction, Transcendence, and Virtual Reality." *Diacritics* 27 (3):77–89.

Wilkes, K. 1994. *Real People: Personal Identity Without Thought Experiments.* Oxford: Clarenden Press.

Wittgenstein, L. (1922) 2001. *Tractatus Logico-Philosophicus*. Translated by D. F. Pears & B. F. McGuinness. New York: Routledge Press.

Wolmark, J. 2002. "Staying with the Body: Narratives of the Posthuman in Contemporary Science Fiction." In *Edging into the Future: Science Fiction and Contemporary Cultural Transformation*, edited by V. Hollinger and J. Gordon. Philadelphia: University of Pennsylvania Press.

Yan, X., B. Zheng, and F. Huang. 2013. "Integrated Motion of Molecular Machines in Supramolecular Polymeric Scaffolds." *Polymer Chemistry* 4 (8): 2395–9.

Young, I. M. 1998. "Five Faces of Oppression." *Philosophical Forum* 19 (4): 270–90.

Zannas, A. S., and A. E. West. 2014. "Epigenetics and the Regulation of Stress Vulnerability and Resilience." *Neuroscience* 264: 157–70.

Journal of
Cognition
and
Neuroethics

Towards an Existentialist Neuroethics

Brandon Fenton
York University

Biography

Brandon D. C. Fenton is a Course Director in Philosophy at York University in Toronto, Canada. His research interests center primarily upon the philosophy of mind/consciousness/cognitive science, philosophy of psychology, action theory, existentialism, and ethics.

Acknowledgements

Presentation of an earlier draft of this work was supported in part by a CUPE 3903 conference travel fund. I would also like to thank Joshua Mugg for his comments on this paper as well as Stefania Mendolia for her comments upon some of the original ideas developed in this paper.

Publication Details

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

Citation

Fenton, Brandon. 2015. "Towards an Existentialist Neuroethics." *Journal of Cognition and Neuroethics* 3 (3): 67–86.

Towards an Existentialist Neuroethics

Brandon Fenton

Abstract

This paper is situated at the intersection of science-fiction literature, existentialist philosophy, and neuroethics—and it amounts to a novel challenge to an implicit neuroessentialism that in large part characterises the field of neuroethics. It begins by examining a character by the name of Peer found in science-fiction writer Greg Egan's well known and award winning novel *Permutation City*. The sub-story of Peer presents a technologically updated image of the Sisyphean metaphor made famous in the existentialist writings of Albert Camus. Although Peer and Sisyphus, in one sense, seem to occupy separate ends on the continuum of freedom, there remains a sense in which both of their stories point to the role of constraints in shaping our ethical agency as well as the inescapability of subjective (moral) choice. Insights drawn from the wide-open imaginative space made possible in the character Peer's malleable virtual world and personal subjectivity lead to a consideration of relevant aspects of ethical subjectivity that are underrepresented (if represented at all) in neuroethical theory. In short, these considerations have to do with ethical subjectivity itself, and the scope of moral freedom. In the end, it is proposed that neuroethical theory be broadened to accommodate concerns about the impact of neuroscientific modifications to the ethical subjectivity of agents. This existentialist turn, while remaining thoroughly natural, eschews an overly simplistic approach to ethical theorizing that is characteristic of reductive neuroessentialism.

Keywords

Neuroethics, Existentialism, Subjectivity, Greg Egan, Neuroessentialism, Sisyphus, Agency

Introduction

Neuroethics, according to Adina Roskies (2002), can be characterized as dealing primarily with two sorts of problems or as having roughly proceeded along two general trajectories. She calls these 'the ethics of neuroscience,' and 'the neuroscience of ethics.' The first includes considerations about whether or not a given neuroscientific research program, in both its design and application, conform to certain ethical standards, as well as an examination of the foreseeable potential legal, ethical, and social impacts of such a proposed study's findings. The other concerns how we may come to understand the operation of traditional ethical notions like value, volition, intention, self-control, freedom of the will, *et cetera*, by way of neuroscientifically studying the functioning of the brain in relevant contexts. This type of research aims to examine, for example, things like how moral values are represented in the brain, whether or not—in terms of brain function—there are any differences between moral and non-moral decision making processes, and in what ways neuronal activity might underlie ethical agency. Later in the same article, following a comment about the openness and appropriateness of the term 'neuroethics' for this burgeoning field of study, Roskies claims that "We should not merely pay lip-service to this inclusiveness. Neuroethics has the potential to be an interdisciplinary field with wide-ranging effects" (23).[1] But despite this proclaimed openness to a variety of disciplines and backgrounds from which to engage and assess both the ways that neuroscientific research projects are developed, implemented, and socially integrated as well as the impacts that such research may have in terms of revisions of traditional ethical notions and theories—all in the service of developing a mature, comprehensive, integrated, and responsible neuroethics—Roskies nevertheless declares that, "Many of us overtly or covertly believe in a kind of 'neuroessentialism,' that our brains define who we are, even more than do our genes" (22).[2] But if we take Roskies to be correct in her characterization of the field, and her observation of one of its central implicit (or often explicit) guiding assumptions (*viz*. neuroessentialism), a tension can be seen to arise between the claim to genuine interdisciplinary openness, and a prevailing sort of reductionism that may threaten to exclude difficult or opposing views right at the very outset.[3]

One of the worries with placing such a neuroessentialist view at the foundation of the neuroethical project is this: if we begin with the assumption that our experience and personal identity can be unproblematically reduced to mere neuronal activity, then we may likewise think that the sorts of ethical deliberations we engage in and judgements we make, as well as the entire edifice of our moral agency can similarly be reduced to the bare mechanistic operation of neurons. We might then even be persuaded to treat value itself as ultimately reducible to the behaviour of neurons.[4] It is in this way that the 'ethics' of 'neuroethics' becomes subordinate to the 'neuro.' One of the problems associated with this sort of view, as Racine notes, is that it appears to "…commit the naturalistic fallacy and threaten[s] to reduce the normative dimension of bioethics [or neuroethics] to biological [or neurological] imperatives" (2010, 55). Another of the

[1] See Racine (2010) for a more recent endorsement and defense of the interdisciplinarity of field.

[2] For proponents of the neuroessentialist view see for example, Gazzaniga (2005), and Churchland (2006). For challenges to the reductive neuroessentialist view see for example, Morse (2006), and Buller (2006).

[3] For more on neuroessentialism, its problems, and its alternatives, see Racine (2010), and Illes (2006).

[4] And this is something that, at the very least, value realists would outright reject (see for example, Nagel's (2012) "Mind & Cosmos").

dangers of allowing this to happen is that we may then be left with an insufficiently nuanced and facile treatment of the role of ethics in this joint area of research. It is my contention that such an unbalanced state of affairs would leave much to be desired when it comes to developing a mature, inclusive, and comprehensive view of neuroethics. And therefore, in this paper, I intend to broaden the ethical scope of neuroethics in a way that challenges the general presumption of neuroessentialist exclusivity and begins to put into practice the sort of broad based inclusivity of various disciplines called for by Roskies. To accomplish this, I will draw upon insights gleaned from science-fiction as well as existentialist literature so as to reveal a lack in current neuroethical theorizing that, when given sufficient attention, allows us to resist the sort of dominating influence of the type of reductionism identified above. Central to this challenge is the notion of the ethical subject and the ways in which neuro-modification or manipulation may impact and undermine the subject *qua* ethical subject.

In terms of layout, the paper contains the following main sections: 1) I will provide some background on Greg Egan's award winning science-fiction novel *Permutation City* and the sub-story of the character named Peer which will serve as my example of the value of the imaginative contributions that sci-fi may present for neuroethics and ethical theory in general; 2) I will relate the example of Peer to the earlier existentialist consideration of the myth of Sisyphus by Camus, as well as present a further analysis rooted in existentialist thought; 3) I will examine what new sorts of existentialist issues we are faced with when we consider how the example connects with neuroethics and the ways in which neuroscience may impact the ethical subject; 4) I will present some concluding comments about how such an existentialist take on ethics resists the sort of reductionism implied by neuroessentialism and why such considerations deserve to be part of the neuroethics discussion.

1. The Irreducibility of Subjectivity in *Permutation City*

Greg Egan's (1994) novel *Permutation City* presents a bifurcated picture of the world in the mid-twenty-first-century. Although the story focuses upon the lives of several fully subjective digital 'copies' of wealthy flesh and blood people who were able to afford entry and are now contained within a virtual reality world,[5] that world is not entirely disconnected from the economic forces still at play within the natural world and the computing power that must be generated there to sustain their virtual existence.

[5] Each of whom, we may presume, would have flatly rejected Nozick's (1974) arguments against plugging into "the experience machine."

This virtual world, like our own, remains economically stratified with the less wealthy copies running at a slower rate than the copies of other more affluent individuals, but the impacts of climate change on the natural world threaten all of the virtual reality world inhabitants the same, since the global computing power upon which they subsist has begun to be diverted by the natural world needs of extreme weather tracking and predicting. This state of affairs provides a snapshot of the background context in which we encounter the sub-story of the character named Peer that I will examine and use as an example from which to draw insights for later arguments and reflections.

One thing to note from the outset, in this story, is that it is clear that Egan adopts a functionalist and reductivist view of not only consciousness, but also subjectivity. In other words, it is not merely conscious thought and experience that is first biologically and then functionally reducible, and therefore, amenable to computational reproduction for Egan, but a subjective sense of self that is able to maintain some sort of integration or unity and continuity that may also be reproduced within the story. The important point about this reductive view of consciousness and subjectivity for Egan, his character Peer, and indeed for us as well, is that it allows for a vision of consciousness and subjectivity that is fully expressible in terms of Turing computability or mechanical relations. That is to say, Peer's virtual-world subjectivity is nothing over and above the mechanistic or computable information-packet transitions that simultaneously constrain and represent it. Indeed, as Farnell (2000) notes, "The reductionist rhetoric of neuro-cyber symbiosis reveals a return to the Cartesian AI notion of 'mind as computation'…that erases the phenomenological model of mind, body, and world" (72). But one of the interesting consequences of adopting this idea as genuinely possible, is that it allows for an imaginative space in which the author (and readers) may explore and examine some of the various ways in which the character Peer can have his subjectivity modified or that he can change and restructure his subjectivity himself—and in the story, as Burnham (2014) notes, this is precisely what he does, by having "…embarked on a grand experiment of self-editing—making it easy to pass time by programming himself to enjoy all sorts of repetitive tasks" (87). Indeed, when we first encounter the virtual reality copy and character named Peer in Egan's story, we find him scaling down from an infinitely tall building towards an ever receding ground. As Egan tells us, "Peer knew he could keep on approaching the ground for as long as he liked, without ever reaching it. Hours, days, centuries" (1994/1998, 60). Peer's being a virtual reality copy in this particular virtual reality world means that he can both design the type of world in which he should want to live as well as the type of attitudes, moods, beliefs, and desires that he should have. In short, although Peer is running at a much slower rate than the more affluent inhabitants of this virtual world, he is nevertheless his own god in a sense—he can create any sort of virtual reality world that he might desire, and he can even manipulate what sorts of desires and satisfactions he will experience within that world. Moreover, the slower refresh rate at which he must operate makes no subjective difference to his private experience as a computational or virtual copy—if he desired to, he could take a snapshot of his cognitive profile at any moment and freeze it for as long as he wished before resuming it without perceiving anything by way of lost subjective time. In fact, his entire cognitive apparatus as well as his subjective experience is completely within his own power to fashion as he sees fit. If he wants to edit out that embarrassing prom night experience that he had prior to becoming a virtual world copy, then he can simply delete that experience from his memory as well as any trace of the impact that such an experience might have had upon his emotional or cognitive states or dispositions.[6] In principle, he could even produce multiple copies of his digital self-consciousness profile to run simultaneously—the concept multiple subjective 'selves' being one that Egan explores here in the character of Paul Durham and in other novels as well.[7]

One of the fascinating things about the character Peer is that, despite this apparent complete freedom to both model his world and actively organize his own thought, mood, emotion, and experience in any conceivable way, Egan nevertheless chose to portray Peer as a modern-day techno-Sisyphus. Contrast this with the image of the original Sisyphean myth in which Sisyphus' fate of rolling a large rock up a hill only to have it roll back down for all eternity—a fate which is commonly taken to be the model of agonizing unfreedom—and a rather striking thematic reversal becomes apparent. But these two Sisyphean views are not only marked by this difference between complete freedom and a total lack thereof, they are also unified under a particular and prominent existential notion about choice. It seems that the earlier existential rendering of the myth of Sisyphus provided by Camus, in which he suggests that the existential challenge of the myth is that "One must imagine Sisyphus as happy" (1942/1988, 111) found a sympathetic ear in Egan who, early on in the story, claims that Peer is in fact "a happy Sisyphus" (61). But before saying anything more about the original myth or its existential analysis, I want to spend some time reflecting on the situation in which Peer finds himself in the story.

[6] There is of course always the looming question with respect to personal identity about how much of one's self can be edited away before one is no longer the same self, but I will leave such questions to the side in this paper.

[7] See for example his (1992) *Quarantine* as well as Hayles (2015) article on that work.

In one sense, the example presented by the sub-story of Peer seems to be situated at one rather extreme end of what we might take to be a spectrum of free agency in that, as mentioned, it more or less renders him a god within his virtual reality world—i.e. due to his ability to entirely craft the world of his own experience and the sort of self that he will have within it as well—whereas the natural world that we all inhabit imposes numerous constraints upon what we may experience and do. However, I don't think that the situation presented in the story is one that is very hard for most of us to at least imagine (which is to say nothing about whether or not we see the example as logically conceivable or metaphysically possible). By now, films like *The Matrix* and other similar science-fiction movies that presume consciousness and subjectivity to be reducible and electronically reproducible have become a part of the landscape of popular culture, and inventions like virtual reality helmets and thought controlled computer interfaces continue to make the fantastical imaginings of yesterday look like the obvious technology of tomorrow.[8] We also know that modifications to our cognitive and physical functioning afforded by modern neurosurgery, neuropsychopharmacology, and other neuroscientific advances have already allowed us to alter our experience of the world in striking ways.[9] So we can imagine being in Peer's virtual shoes, so to speak. This is why it is so curious that Egan chose to fashion Peer as a sort of Sisyphean character. Given that most readers could fairly easily accept the speculative ideas being made use of in the novel, and would likely want to explore far more exciting experiences in such an open landscape if granted the same sort of opportunity, readers are left to wonder why Egan opted to make Peer the image of repetitive drudgery. Perhaps Egan thought that repetitive activity was essential to maintaining some sense of connection to the prior flesh and blood human that the digital copy Peer once was—or at least, believed himself to be. Maybe we the readers of this story would struggle to identify with such a fantastically set subjectivity if it strayed too greatly from our own everyday sorts of subjective experience. But another potential reason for his opting to do so is that Egan recognised that the existential perspectives and questions of life will remain in any post-human future insofar as there exists some form of subjectivity or self-consciousness. As Heidegger suggests, our personal "Being is that which is an issue for every such entity" (1927/2008, 67). In other words, regardless of the context in which subjectivity manifests—be it organically or digitally—one's subjectivity is always a central concern or problem that a subject faces simply in virtue of being a

[8] Granted, something like virtual consciousness or subjectivity still appears to be a rather far off dream.

[9] See for example, Crockett *et al.*, 2015.

subject. Indeed, and this in part because, as Sartre claims, "Every conscious existence exists as consciousness of existing" (1943/1984, 13). For Peer, as for the rest of us, we must each ask ourselves what it is that makes our lives meaningful and worth living. And this is a question that the existentialists recognize we must all answer for ourselves.

The story of Peer is crafted such that, even with a virtually unlimited degree of freedom in which to shape himself and his environment—as well as what a given environment will mean to him once experienced from the inside—he nevertheless decides to adopt a Sisyphean life of consistent physical exertion at a single basic activity (i.e., scaling the building infinitely). We might think: how very human of the copy Peer to constrain his activity in this way. Let us not forget that Peer has complete authority over how he might feel or think about any of this—there is no danger that he will grow bored of this activity, such a possibility has been edited out of the cognitive script that he chose to adopt for himself. Likewise, there is no danger that old memories may interfere and distract him or lure him from his activity with the promise of something better or at least something different—the cognitive structures or patterns of activity that represent these too have been sectioned off from his self-selected model of himself. Indeed, he knows only how to be happy with the project that he has selected for himself regardless of what any of us may think of it. One of the salient features of Peer's paradise (as we might be wont to call it) is that the only constraints that he experiences are those that he has imposed upon himself. And those self-selected constraints are the only markers by which we can identify Peer as, in some way, human, or as the digital descendant of a human that retains something of its former flesh and blood self—even if that is now little more than a highly plastic, digital rendering of a particular neural architecture and its general activation patterns. Later on, I will have more to say about how it is that such constraints condition our experience of the world and shape our ethical subjectivity within it. Next, however, I would like to take a moment to consider Camus' existentialist understanding of the original myth of Sisyphus before examining how it connects with the story of Peer.

2. Camus' Sisyphus

As mentioned earlier, the original mythical story of Sisyphus is one of a man condemned by the gods to push a boulder up a hill only to have it roll back down to the bottom over and over again, for all time. According to Camus, the story of Sisyphus is standardly conceived of as the mythical metaphor of a repetitive, toiling, and apparently meaningless life. Indeed, he claims: "Sisyphus is the absurd hero. He is, as much through his passions as through his torture. His scorn of the gods, his hatred of death, and his passion for life won him that unspeakable penalty in which the whole being is exerted toward accomplishing nothing" (1942/1988, 76). However, although Camus characterizes the fate of Sisyphus as being absurd, he nevertheless sees something heroic in Sisyphus that seems to be overlooked by the casual observer of the story. Pondering that pause between his having just rolled the rock up to the top of the hill and having to turn and retrieve it again from the bottom, Camus says:

> That hour like a breathing-space which returns as surely as his suffering, that is the hour of consciousness. At each of those moments when he leaves the heights and gradually sinks toward the lairs of the gods, he is superior to his fate. He is stronger than his rock. (76)

And it is the subjective sentiment of scorn that reveals the heroic strength of the Sisyphean love for life and hatred of death. Again, in the words of Camus, "The lucidity that was to constitute his torture at the same time crowns his victory. There is no fate that cannot be surmounted by scorn" (77). It is in this way that Camus characterizes the existential triumph of Sisyphus; who is at once driven by scorn to both defy the punishment of the gods and to overcome his fate by, in a sense, 'owning' that very fate and finding the joy of his subjectivity therein. The one thing that not even the gods have dominion over is his very subjectivity. As Camus says "His fate belongs to him. His rock is his thing" (78). This personal subjective recognition of one's life, in any form that it may take, is central to the existential perspective. Indeed, Sartre claims, in one of his most famous lectures on existentialism, that "As our point of departure there can be no other truth than this: *I think therefore I am*. This is the absolute truth of consciousness confronting itself" (1946/2007, 40).[10] It is a subjectivity that remains non-reducible because it is that which is ultimately free and that by which we may come to understand objects in the first place. This sort of radical freedom is also at the heart of an existentialist approach to ethics. Indeed, with respect to ethics, Sartre suggests that, the existentialist "…can will but one thing: freedom as the foundation of all values" (1946/2007, 48). Mirroring this view of the centrality of the importance of an ultimately free subjective choice is the personal perspective that Sisyphus adopts towards his fate in Camus' retelling of the story. In that version of the myth, Sisyphus' subjective acceptance

[10] While this quote provides a rough and ready notion of Sartre's view of the nature of self-consciousness, he develops a much more thorough account in his (1943/1984) *Being and Nothingness* (see especially section 3 of the introduction).

of his existence is as freely chosen as the perspective that he may adopt toward any sort of life. And it is this same feature of subjective irreducibility that we find reflected in the character of Peer. On the one hand, we have Sisyphus, the model of the unfree labourer, who nevertheless triumphs over his fate by way of the freedom of his subjectivity; on the other, we have Peer, the model of absolute (or next to divine) freedom who happily chooses to narrow his activity to a single subjective project. In both cases, however, (i.e. the apparently unfree and the seemingly absolutely free) there remains a power to decide that, although conditioned by various constraints—in one case natural, in the other due to computing power—is not entirely constituted nor caused by them. This is one way in which the reductivist rendering of things in Egan's story might be seen to begin to unravel—it is one question just how much Peer may modify his cognition while still remaining Peer;[11] it is another to inquire into the difference between Peer and a program that performs the same functions while yet not amounting to a subjective being. I am concerned with this latter question. If the virtual world Peer is a genuinely self-conscious subject, his subjectivity is an issue for him. It is something that belongs to him as such a being and his choices must be made in light of being a subjective being. On the other hand, if the program that represents Peer is merely running through various transformations of digitally encoded information over time, then it is at best only *subject to* such transformations and never the *subject of* them. That is to say, *that sort of Peer* entirely lacks such subjective choice.

3. Existential Implications for Neuroethics

What is perhaps most compelling about the example of Peer and the existentialist lens through which we can interpret the story, is how it gives shape to what we may call the ethical subject[12], and how changes to the ethical subject matter to neuroethics. I see the notion of the ethical subject as, in a sense, partially falling in between what Roskies categorized as the "ethics of neuroscience" and "the neuroscience of ethics." As mentioned, for Roskies, the ethics of neuroscience is concerned with "the ethical issues and considerations that should be raised in the course of designing and executing neuroscientific studies and [an] evaluation of the ethical and social impact that the

[11] This question of the limits of modification and personal identity is raised at the end of Egan's novel in the character of Paul Durham (1994/1998, 307), and is examined further in Farnell (2000).

[12] My understanding of the notion of the ethical subject is in large part congruent with Simon Critchley's (2012) proposal but I will not elaborate on what is entailed by that view here. See his entry in references for further clarification.

results of those studies might have or ought to have on existing social, ethical, and legal structures" (2002, 21). Whereas she sees the neuroscience of ethics as the investigation of traditional ethical notions such as free-will, self-control, personal identity, intentionality *et cetera*, in terms of brain functions. She maintains that the neuroscience of ethics can be framed in terms of questions like: "How are decisions made in the brain?" and "How are ethical decisions similar or different from other types of decisions?" (2002, 22). As noted, this latter approach is, Roskies admits, if not explicitly, then at least typically implicitly sustained by a sort of reductive neuroessentialism, or the view that it is in fact our brains that entirely determine the choices that we make and the sorts of persons that we are. And this is one space in which I think that the example of Peer and the existentialist perspective has something to contribute to the project of neuroethics—if not by directly challenging certain fundamental assumptions of the field, then at least by cautioning the discipline against an overly simple way of approaching ethics.

Adopting an existentialist perspective when considering the standard terrain of neuroethics certainly problematizes things, but it also affords us an opportunity to re-examine certain basic commitments and assumptions and to identify certain subtle concerns that may otherwise be overlooked. With respect to Roskies' first category of the 'ethics of neuroscience', the existentialist view (as I will refer to it)[13] reminds us here that ethical actions are not simply a matter of plotting the costs and benefits of some neuroscientific study against the predefined structures of a deontological, or utilitarian, or virtue ethical list of do's and don'ts. Instead, genuine and authentic moral behaviour is something chosen by an engaged subject who is responsible for the selected behaviour. As something subjectively and irreducibly chosen, ethical behaviour cannot be entirely captured calculatively and mechanically—this reminds the researcher, for instance, that, as a subject herself, she remains responsible for the types of projects that she decides to undertake regardless of the operational norms of the discipline or society at large, and that ethical action is about more than the mere application of and adherence to a given codified list of prescriptions and proscriptions. Indeed, it remains always, first and foremost, a responding to the ethical demand *by* and *as* a subject.

[13] By my use of the phrase 'the existentialist view' I do not mean to imply that my particular reading of existentialist literature is perfectly doctrinaire or that there is a single existentialist view to be appealed to. Rather, my take on the existentialist view presented in this paper reflects something of the widely examined dominant themes of much existentialist literature; themes like radical freedom, subjectivity, thrownness, *et cetera*.

In terms of Roskies second category of 'the neuroscience of ethics', the existentialist view appears to stand in direct conflict with the reductive 'neuroessentialism' of this approach. But it should be mentioned here that the existentialist view does not necessarily deny the hard facts of the world (or of science)—instead, it reminds us that even such facts are first interpreted by a subject and thus, our understanding of our own subjective decisions are at least on par with the determinations of the sciences.[14] But I don't now intend to defend the existentialist view from a form of reductive materialism. Instead, I want to use the example of Peer, cast in a certain existentialist light to draw attention to a perspective that I take to be relevant to—and commonly overlooked by—neuroethical theorizing.

There are two central aspects of an existentialist view of ethics that I want to highlight. First, is the notion of freedom as one of the primary and yet ungrounded values of the existentialist view[15]; second, is the notion that the ethical context is one in which the ethical subject is responsive to and experiences a certain ethical demand. This ethical demand can also be characterized in terms of something making a claim upon the subject or the subject experiencing a particular type of behavioural constraint.

Imagine, for example, that you encounter a person physically harming a child. In this situation, it is the child's defencelessness, and experienced harm that calls on you to intervene and put an end to the abuse. Another way of thinking of this sort of situation is to frame it in terms of ethical constraints. You remain free to either respond ethically and intervene, in order to stop the abuse, or you may also choose to ignore the child's plea and carry on with your own affairs—failure to respond here being something that you are responsible for, and something that merits reproach or moral condemnation. The ethical constraint presents itself to you (the ethical subject) as a demand or request for intervention and authentic engagement in the moment; regardless of whatever ethical system you might generally endorse (if any). The experiential landscapes of our ethical lives are constrained by innumerable such ethical demands by others (some much more benign, and some even more troubling). The homeless person who asks: "Will you provide me with something to eat?" The oppressed peoples who ask: "Will you protect us from further violence?" The worker who asks: "Will you pay me a living wage that I may

[14] It is also important to note that subjectivity and personal agency, from an existentialist point of view, do not require any form of supernatural or substance dualist intervention in the natural world.

[15] The importance of the notion of radical freedom to the existentialist works of Jean-Paul Sartre, for example, can hardly be overstated—and central to that notion is the view that "subjectivity must be our point of departure" (1946/2007, 20).

care for my children?" These and many other demands extend beyond the personal to the social, the ecological, and other domains: "Will you stand with the people, for economic, social, and political equality?"; "Will you protect wildlife from extinction?"; "Will you act to spare the next generation from the consequences of climate change?" and so on.

But let us now return to the example of Peer. As mentioned, Peer lives in a world of his own creating and experiences a subjectivity that is constrained in a self-selected way. This apparent absolute freedom to self-organize and to re-organize self may present itself as a post-human fantasy but it has a clear implication for how we are to understand what it is to live ethically. Although Peer's virtual world activity is constrained to something all too human (*viz*. a repetitive pattern of physical behaviour), in his virtual world, he faces none of the ethical demands that we regularly encounter in the natural world. There is no environmental constraint the likes of which calls upon him to act in one way over another. The constraints under which he lives are merely procedural, and they affect no one other than himself. The fact that his solipsistic existence is connected to a larger natural world that is suffering various economic and ecological crises is something that Peer has simply 'edited out' of his cognition. But this sort of editing out of larger experience is a serious ethical worry that carries over into the more modest interventions of modern day neurosurgery and neuroscientific modifications of cognitive functioning. Within his solipsistic world, Peer appears not as immoral but rather, simply amoral—i.e. the notion of ethical conduct simply doesn't seem to apply to the sort of being that Peer supposedly is, in the sort of world in which he resides. However, if we take the broader perspective of his absence from the natural world into account, it becomes apparent that his opting to retreat from the ethical demands of his time and place in the natural world to be a complete abnegation of his ethical responsibility—and insofar as his restructuring of his digital neuro-architecture is aimed at eliminating his *freedom to respond* to the ethical demands of the larger world, it too is deeply immoral. It is immoral both in the sense that it restricts his ability to respond to various ethical demands and in the sense that it destroys the scope of his very subjectivity—the former amounting to a limitation on the social or relational aspect of his ability to respond ethically, and the latter being a limitation on the sort or ethical subject that Peer could otherwise be. And to me, this sort of minimizing of the scope of one's ethical subjectivity is already a problematic feature of the way in which human beings modify their cognitive functioning—either by way of neuropsychopharmacology, neurosurgery, or otherwise—that neuroethics ought to be both cognizant of and engage with more substantively.

Allow me to illustrate the worry as I see it. I may, for instance, be depressed and distressed by having, for example, witnessed the unjust and violent oppression of a given group of people by the state, but taking a little blue pill will effectively modify my brain function such as to alleviate my depression and leave me feeling unmoved by such concerns. I might likewise feel anxious about confronting a misogynist employer regarding his treatment of women workers, but some other neuro-chemical fix might permit me to look the other way with minimal discomfort, and so on. But such modifications to my subjectivity take something important away from me. These interventions remove from me my ability to be fully present, and engaged by the ethical constraints that the world presents me with. Indeed, this sort of "cosmetic pharmacology" as Peter Kramer (1993/1997) dubs it, alters in a deep and abiding way the very ethical subject that I am or that I would otherwise be, warts and all. By decreasing or eliminating my ability to be sensitive and receptive to the ethical constraints or demands of regular life, such modifications undermine my subjectivity and my freedom to become the kind of ethical agent I might otherwise have the chance to be. Therefore, I see the task of neuroethics not only as, for example, identifying those operations or modifications of the subject that are unethical because they come at too great a risk or cost to a particular patient, or society, or to some other dimension of the patient's quality of life or what have you; but also as coming to terms with the more subtle ways in which treatments, therapies, and cognitive modifications may function to undermine or excise portions of the agent's very ethical subjectivity itself in ways that may result in a *narrower sensitivity* to the ethical demands that the agent is presented with in the world.

But reflection upon such considerations does not always present careful researchers with obvious answers. Take, for instance, the following example: an American veteran of the Iraq war suffers from post-traumatic stress disorder (PTSD) as a consequence of his having survived a road side bomb attack which killed several of his fellow soldiers during his tour of duty. The personality changes and anxiety attacks that result from his PTSD while both deeply impeding his ability to function well socially and in the civilian workforce nevertheless provide him with the impetus to reflect upon the horrors of war and to commit to writing a memoir that exposes some of the atrocities in which he had taken part as an ethically motivated gesture of atonement.

On the one hand, you have the soldier's anxiety and personality issues which are causing trouble for him in his daily social interactions and work life. And here it seems that any neuropsychopharmacological or other neuroscientific treatment that enables the soldier to better navigate his day to day life is to be desired. The apparent benefits here being that he may both no longer suffer from the haunting images of his experiences in the war (or at least have to deal with these flashbacks much less frequently), and he may begin to do better in his social and work life. But on other hand, the psychological consequences of his lived experience of the war, if untreated,[16] would lead to an act of ethical agency—i.e. the writing of the exposé/memoir as an ethical gesture of atonement. It seems to me that in most, if not all cases, the soldier would likely be given whatever sort of treatment is available to improve his quality of life and social functioning; and that this would be assumed to be an ethical way of helping to treat an individual who is struggling with the reality of his lived experience. The problem remains, however, that this sort of narrow approach to what it means to reflect on things ethically entirely misses the point about the ethical subjectivity of the soldier himself. If such treatment dulls the soldier to his memories and lived experience in such a way that it restricts the scope of his ethically responding to demands that he would otherwise answer, then it does him a disservice and impacts his ethical subjectivity in a way that is harmful as well.

I don't have a clear answer as to what ought to be done in such cases—i.e. whether we ought to value the soldier's peace of mind and social integration above his neurochemically unaltered ethical subjectivity—but what I am arguing is that alterations to the scope of his ethical subjectivity deserves far greater consideration than it appears to typically receive in neuroethical theorizing.[17] One of the reasons that considerations about the ethical subjectivity of a patient or research subject might not be as prevalent in the literature may have to do with, as suggested earlier, its partially falling in between the two standard research categories identified by Roskies. Indeed, while 'the ethics of neuroscience' might provide us with guidance when it comes to how to avoid the obviously socially harmful, legally objectionable, or other reductions to the quality of life of a given patient, it appears to overlook questions about the ethical subjectivity of a patient or research participant because the focus tends to be more squarely set upon the discipline of neuroscience as an ethically accountable practice or metaphorical ethical agent in its own right. And to the extent that 'the neuroscience of ethics' aims to ultimately reduce ethical notions to more basic neural processes, it fails to acknowledge that subjects respond to ethical demands first and foremost *as* conscious subjects. So it seems clear that an existentialist understanding of ethical subjectivity amounts to, if not

[16] We will presume for the sake of argument that it will only be in the case of not receiving treatment that the soldier is motivated by his PTSD symptoms to write the exposé/memoir.

[17] So far we have only been considering neuroscientific modifications to brain function that are presumed to limit or reduce the scope of one's ethical subjectivity but we might also argue about whether or not modifications that enlarge the scope of one's ethical subjectivity (by making one more sensitive to ethical demands that one might normally fail to notice) ought to be pursued. However, I will save my thoughts on arguments about the prospects for an enlarged scope of ethical subjectivity for a future paper.

an alternative to standard approaches to neuroethics, then at least a corrective to an over-simple view of what ethics might entail.

4. Existentialist neuroethics and neuroessentialism

As mentioned previously, one of the two main branches of neuroethics—the one that Roskies calls 'the neuroscience of ethics'—is often characterized by a commitment to neuroessentialism, or the view that it is the functioning of the brain that entirely determines the types of people that we are, as well as the sorts of ethical behaviours which we will perform in various circumstances. In direct contrast to this strongly neuro-deterministic view lies an existentialist understanding of ethical subjectivity that takes radical freedom to decide and personal subjectivity as the starting point of any realistic account of ethical agency. Clearly, these two positions appear to be at odds with one another. And there appears to be a problem with attempting to maintain that these views are in any way compatible. The point has been made by Žižek (2010/2011) that transhumanists often fail to see this sort of issue even as it stares them in the face:

> …when they describe the possibility of intervening in our biogenetic base and changing our very "nature," they somehow presuppose that the autonomous subject freely deciding on his or her acts will still be present, deciding on how to change its "nature."…on the one hand, as the object of my interventions, I am a biological mechanism whose properties, including mental ones, can be manipulated; on the other hand, I (act as if) I am somehow exempt from this manipulation, an autonomous individual who, acting at a distance, can make the right choices. But what…[if]…the autonomous individual is no longer there? (347)

In other words, the contrast in views appears to be insurmountable. Either we accept the neuro-essentialist assumption that we are thoroughly determined by our brains, or we assume that we are radically free in a way that neuroscience could never alter nor impair because it deals only with neurons and not subjects of experience. The astute reader will have noticed that this tension between the strong determinism of neuroessentialism and the radical freedom of existentialist subjectivity has been in the background of this paper for almost the entire time—but I have not made the mistake with which Žižek charges the transhumanists since my argument is that the subject in fact is altered by neuroscientific modifications *to his or her being*. However, I don't think that this means that neuroessentialism therefore comes out on top, and I don't think that things are quite as black and white as the above quote frames them. Indeed, I think we can find a middle way between these two apparently opposing views if we simply soften the edges of each—and insofar as we are aiming at an inclusive and comprehensive approach to neuroethics, doing just so looks to be a worthwhile objective.[18]

In order to avoid the apparent tension between the two identified approaches we need only understand the way in which they might work together in a sort of hybrid form. To accomplish this we might acknowledge the *influence* of neuro-modification upon the scope of the ethical agent's subjectivity by affecting her moods, affect, attention, attitudes, *et cetera*, while maintaining that such an influence does not utterly determine—in other words, only partially constrains—the final choices of the ethical subject, since such choice is only sensible to the ethical subject *qua* self-conscious subject. That is to say, while the subject's choice can remain ultimately free, the range of things over which she may be consciously aware can be restricted or impacted by neuroscientific interventions just as they can by other physical interventions. Additionally, we will need to soften the notion of radical freedom that is at play in the existentialist view as well in order to make room for the fact that subjective choice can be impaired by limiting the scope of things to which an agent remains receptive or cognizant. Yes, there may be a sense in which one's subjectivity and choice remain ultimately free, but if one is kept from developing an awareness of certain things due to neuroscientific interventions, then the scope of one's freedom is impaired just as much as one's movement is compromised by being stuck on an island and not knowing how to swim.

Ethics arises in a context of constraint; in a context of a demand that is experienced by the ethical subject—any neural modification that diminishes the ethical subject's sensitivity to the natural ethical demands of the world harms both the ethical subject or agent as well as those sources of ethical demands whose call for concern goes unanswered. And any overzealous attempt to completely reduce ethical agency (or the ethical enterprise itself) to neuronal happenings fails to understand the finer points of ethical reflection and action as well as drastically over-estimates the kinds of things that neuroscience can tell us. But there is reason to be hopeful that we can avoid these types of errors in the future once they are more widely recognized and acknowledged. As Parens & Johnston (2007), suggest:

[18] This more modest 'middle way' that I am suggesting here is largely consistent with what Racine (2010, 65) calls a 'moderate pragmatic naturalism.'

> It might indeed be possible for neuroethicists to work closely with neuroscientists without succumbing to the hyperbole that genethicists once succumbed to at the elbows of geneticists. As we work to resist that temptation, we need to be vigilant about using the complexity-reducing shorthand that scientists, journalists, bioethicists and others often use. When we hear anyone talk of 'the part of the brain for' complex behaviour X, we should remember that, once upon a time, geneticists spoke of 'the genes for' complex behaviour X. (S62-S63)

So rather than falling prey to the inadequacies and exaggerated promises of a neuroessentialist perspective, let us neuroethicists increase the scope of our ethical reflections to include consideration of the ethical subject and how neuroscientific interventions might impact the very subjectivity of ethical agents by impeding their freedom to respond to the sorts of ethical demands that everyday life presents to them.

References

Buller, T. 2006. “Brains, Lies, and Psychological Explanations.” In *Neuroethics: Defining the issues in theory, practice, and policy*, edited by J. Illes, 51–60. New York: Oxford University Press.

Burnham, K. 2014. *Greg Egan*. Modern Masters of Science Fiction. Chicago: University of Illinois Press.

Camus, A. (1942) 1988. *The Myth of Sisyphus & Other Essays*. London: Penguin.

Churchland. S. 2006. “Moral Decision-making and the Brain.” In *Neuroethics: Defining the Issues in Theory, Practice, and Policy*, edited by J. Illes, 3–16. New York: Oxford University Press.

Critchley, S. (2012). *Infinitely Demanding: Ethics of Commitment, Politics of Resistance*. London: Verso.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal O.T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. 2015. “Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making.” *Current Biology* 25: 1–8.

Egan, G. (1994) 1998. *Permutation City*. London: Millennium.

Egan, G. 1992. *Quarantine*. New York: HarperPrism.

Farnell, R. 2000. “Attempting Immortality: AI, A-Life, and the Posthuman in Greg Egan’s *Permutation City*.” *Science Fiction Studies* 27 (1): 69–91.

Gazzaniga, M. S. 2005. *The Ethical Brain*. New York: Dana Press.

Hayles, N. K. 2015. “Greg Egan’s Quarantine and Teranesia: Contributions to the Millennial Reassessment of Consciousness and the Cognitive Nonconscious.” *Science Fiction Studies* 42 (1): 56–77.

Heidegger, M. (1927) 2008. *Being and Time*. New York: Harper Perennial.

Illes, J. 2006. *Neuroethics: Defining the Issues in Theory, Practice, and Policy.* New York: Oxford University Press.

Kramer, D. (1993) 1997. *Listening to Prozac: A Psychiatrist Explores Antidepressant Drugs and the Remaking of the Self*. Revised edition. New York: Penguin.

Morse, S. J. 2006. “Moral and Legal Responsibility and the New Neuroscience.” In *Neuroethics: Defining the Issues in Theory, Practice, and Policy*, edited by J. Illes, 33–50. New York: Oxford University Press.

Nagel, T. 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. New York: Oxford University Press.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

Parens, E. & Johnston, J. 2007. "Does it Make Sense to Speak of Neuroethics? Three Problems with Keying Ethics to Hot New Science and Technology." *EMBO Reports* 8: S61–S64.

Racine, E. 2010. *Pragmaitc Neuroethics: Improving the Treatment and Understanding of the Mind-brain*. Cambridge: The MIT Press.

Roskies, A. 2002. "Neuroethics for the New Millenium." *Neuron* 35: 21–23.

Sartre, J-P. (1946) 2007. *Existentialism is a Humanism*. Connecticut: Yale.

Sartre, J-P. (1943) 1984. *Being and Nothingness*. New York: Washington Square Press.

Žižek, S. 2011. *Living in the End Times*. 2nd edition. London: Verso.

Journal of
Cognition
and
Neuroethics

Two Minded Creatures and Dual-Process Theory

Joshua Mugg
Indiana University Kokomo

Biography

Mugg works on the cognitive architecture of human reasoning, the nature and ethics of belief, and the import of these two areas for philosophy of religion. He did his graduate work at York University in Canada, and is now a Visiting Lecturer at Indiana University Kokomo.

Acknowledgements

Many thanks to Brandon Fenton for comments on an early abstract and (nearly) final version of this paper. Thanks to the audience and organizers of the first Cognition and Neuroethics in Science Fiction Conference.

Publication Details

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

Citation

Mugg, Joshua. 2015. “Two Minded Creatures and Dual-Process Theory.” *Journal of Cognition and Neuroethics* 3 (3): 87–112.

Two Minded Creatures and Dual-Process Theory

Joshua Mugg

Abstract

How many minds do you have? If you are a normal human, I think only one, but a number of dual-process theorists have disagreed. As an explanation of human irrationality, they divide human reasoning into two: Type-1 is fast, associative, and automatic, while Type-2 is slow, rule-based, and effortful. Some go further in arguing that these reasoning processes constitute (or are partly constitutive of) two minds. In this paper, I use the *Star Trek* 'Trill' species to illuminate the condition for the existence of "two minds in one brain" (Evans 2010, 3). After carefully outlining the two dominate versions of dual-process theory (default-interventionism, espoused by Evans, Stanovich, and Kahneman, and parallel-competitive theory, espoused by Sloman, Frankish, and Carruthers) and contrasting each with a one-system alternative, I argue that these three views should be understood as existing on a continuum: there are some theories that could plausibly be characterized as either one-system or default-interventionist, and the distinction between default-interventionism and parallel-competitive theory is not as clean-cut as usually assumed. I then argue, using the conceptual claims I defended using the science fiction cases, that default-interventionist dual-process theory is not compatible with the claim that humans have two minds (contra Evans and Stanovich).

Keywords

Dual-Process Theory, Two-System Theory, Two-Mind Theory, Cognitive Architecture, Human Reasoning, Belief

1. Introduction

A reoccurring strategy for explaining irrationality is that of dividing the mind into separate parts. This strategy goes back at least as far as Plato, who, in the *Republic*, argued that the soul is divided into reason, desire, and appetite because "the same thing will not…undergo opposites in the same part of itself, in relation to the same thing, at the same time" (436c, Grube translation).[1] Dual-process theory is the latest iteration of this strategy. According to the Standard View of dual-process theory, reasoning problems cue two very different kinds of processes: Type-1 processes are fast, evolutionarily old, associative (or heuristic), and automatic, while Type-2 processes are slow, evolutionarily new, rule-based, and controlled (or effortful). Some theorists argue Type-1 and Type-2 processing are carried out by two different kinds of systems, System 1 and System

[1] I will not attempt to trace the various iterations of this strategy throughout western philosophy (for an overview see Frankish & Evans, 2009).

2, respectively (Carruthers, 2009, 2013; De Neys, 2006, 2012; Kahneman & Frederick, 2002; Sloman, 1996, 2014). Others are agnostic as to how many systems there are and emphasize only Type-1 and Type-2 processing to the exclusion of System 1 and System 2 (Evans, 2008, 2009; Evans & Stanovich, 2013).

The most radical version of dual-process theory has it that humans possess two *minds,* one corresponding to Type-1 or System 1, the other to Type-2 or System 2. As Evans puts it, humans have, "in effect, two minds in one brain" (2010, 3, see also Frankish, 2004; Stanovich, 2011). Furthermore, these theorists are not referring to split-brain patients or subjects with multiple-personalities; they are making a claim about normal adult humans. Frankish (2010) claims that "if our judgments and actions" are generated by one of two distinct mental systems, "then many traditional philosophical questions will need to be recast to allow for this duality, with implications for debates about agency, autonomy, responsibility, rationality and knowledge, among other topics," adding that this is "likely to be fertile area for future research" (923. For examples to do just this, see Fiala, Arico, & Nichols, 2011; Mallon & Nichols, 2011; Nagel, 2011). Understandably, dual-process, two-system, and two-mind theories are not without their opponents (see Keren & Schul, 2009; Kruglanski, 2013; Kruglanski & Gigerenzer, 2011; Mugg, forthcoming, 2013; Osman, 2004).

My purpose in this article is to assess the relation between the two dominant versions of dual-process theory and the two-mind theory, arguing that one of these versions (default-interventionism) is incompatible with two-mind theory. To do so, I first use the *Star Trek* 'Trill' species to illuminate the conditions for the existence of two minds in one brain (Section 2). Examining science fiction examples offers a method of examining conceptual possibilities, and offers how-possible models. That is, science fiction helps us understand what the structure of human cognition *might be*. Conceivability is our guide to possibility, since a state of affairs is possible if and only if it contains no contradictions. However, we must conceive in a maximally possible way to check for contradictions. Doing so is difficult. Certain genres of fiction, namely those that do not loosen conceptual constraints, can be fruitful in aiding our conceiving in a maximal way. Of course, the actual nature of human cognition is a matter for empirical investigation, but having how-possible models illuminates how to empirically investigate the structure of human cognition. In section 3, I turn to three empirically motivated positions of the cognitive architecture of human reasoning: parallel dual-process theory (according to which two reasoning systems operate at the same time and in direct competition with one another), default-interventionist dual-process theory (according to which one system is the default, but can be overridden by a second system), and one-system theory (according to which there is one reasoning system that operates in many modes). Against the standard way of thinking about these accounts, I argue that we may see these three views on a continuum, and that there may be borderline cases, especially between one-system and default-interventionist theories—the more integrated the two systems are, the less plausible it is that they are genuinely distinct systems. I then apply the conditions gleaned from science fiction in Section 2 to these dual-process theories (Section 4), arguing that default-interventionist dual-process theory is not compatible with the stronger two-mind theory (contra Evans and Stanovich).

2. Distinguishing Systems, Distinguishing Minds

In this section I will outline the conditions on humans possessing two minds using the *Star Trek* 'Trill' species.[2] Let me start with a few details about the Trill. They are a humanoid species very much like humans. However, a small percentage of the species are unique: they are 'joined' with a symbiont. While a humanoid Trill's natural lifetime is about the same as that of a human, the symbiont's is much longer, and the symbiont is passed from host to host. Each new host gains all the memories, experiences, and even (to some degree) personality traits of the former hosts. This is why, when Jadzia-Dax, a joined Trill (Jadzia is the host, Dax the symbiont), has her symbiont stolen, she says that she "feels so empty" ('Invasive Procedures'). The symbiont, under normal conditions, is integrated into the Trill's nervous system. Indeed, after the two are joined for 94 hours, it would kill the host if the symbiont were removed for longer than a few hours (see 'Dax' and 'Invasive Procedures'). Thus, the symbiont and host are two biological systems that can operate fairly independently. Dr. Bashir says they are "like two computers linked together" ('Dax'), but, under normal conditions, they depend on each other in important ways: the symbiont is dependent on the host for nutrition and life support, and (after being joined) the host is dependent on the symbiont to continue living (though Jadzia can survive for a short time without Dax).

Two distinct humans have distinct minds. The fact that they are distinct, biological, minded creatures is sufficient for their minds to be not identical. I will call this the Organism condition:

[2] I will focus on the *Deep Space Nine* version of the Trill, in which the resulting host and symbiont are a blend. However, I must note that in the original appearance of the Trill, in *The Next Generation* 'The Host,' the resulting symbiont-and-host aggregate's personality was entirely that of the symbiont. The relation between the two was more like the relation between body and mind on Platonic dualism (at least in the *Phadeo*).

> **Organism Condition:** If X and Y are minded organisms and X is not identical to Y, then X and Y have distinct minds.

The above conditional should not be replaced with a bi-conditional, since, if being distinct organisms were a necessary condition on two-mind theory, then it would be metaphysically impossible that humans possess two minds. It is at least metaphysically possible that humans possess two minds. Thus, I will leave the Organism Condition as a conditional.

Now, Jadzia-Dax is not a single organism. She is an aggregate of two organisms: Jadzia, the humanoid Trill, and Dax, the symbiont. In 'Dax', an arbitrator must decide if Jadzia-Dax is the same person as Curzon-Dax. Toward the end of the episode, she tells Jadzia-Dax "You are either 200 years older than I am, or you are about the same age as my great-granddaughter. At first I wondered which of those you were, now I am bothered by the likelihood that you may be both." This idea is reinforced throughout the series: Jadzia-Dax are two distinct creatures linked together.

The fact that Jadzia and Dax are distinct creatures—indeed, members of different species—is sufficient for them to possess distinct minds. On its own, the Organism Condition does not shed much light on the dialectic between one and two-mind theory, since the putative two minds in the human case would belong to a single organism. However, the Organism Condition does establish that Jadzia-Dax has two minds, and can assuage some immediate worries about the two-mind theory.

First, one might object to Jadzia-Dax's having two minds on grounds that all of Dax's behavior is mediated through Jadzia's body. In reply, notice that Jadzia and Dax's minds are dissociable. Dax can, and eventually does, live in another body. Also, when Dax is taken away from Jadzia, Jadzia can talk, think, and reason. Granted, she is not able to do this for long, since a joined Trill will die without her symbiont, but the point is that Jadzia's cognition can continue (for a time) without Dax's. Thus, they are dissociable. Indeed, Jadzia and Dax's minds are doubly dissociable. The implication for the two-mind theorist is that it is a conceptual possibility that two minds could share one body (though in the human case the minds must be distinguished in some way other than the Organism Condition).

Second, one might object to Jadzia-Dax's having two minds on grounds that Jadzia and Dax are too neurologically integrated to have two minds. Jadzia feels Dax's pain and *vice versa*. However, notice that, in some cases of joined twins, the two children can feel what is happening to the other's body. However, the joined twins have distinct minds. One might object to my counterexample by claiming that joined twins have only one body. I am not convinced that joined twins have one body (in the relevant sense), but suppose my interlocutor is right. If so, we have an instance of two minds (that are somewhat integrated) existing in one body. Thus, replying that the joined twins share a body actually supports my claim that Jadzia-Dax has two minds. The upshot is that it is a conceptual possibility that humans have two minds in one brain.

Jadzia-Dax has two minds. What are some further ways we might be able to tell that she has two minds? In one episode, 'Equilibrium', Jadzia-Dax discovers that one Trill, who formerly possessed Dax, named Joran-Dax, has long been repressed in Dax's cognition. Joran was a violent man, and murdered the previous possessor of Dax (Torias-Dax). Jadzia did not have access to the information that Dax had been involved in a murder. However, Dax did have access. Inaccessibility is not sufficient for distinguishing distinct *minds* (though it may be sufficient for distinguishing cognitive systems, especially modules). I do not have direct access to the process by which I see the screen I am currently looking at, but my lack of access does not imply that the perceptual process is not part of my mind. Something further is going on in the case of Jadzia and Dax's access to information concerning Joran. I suggest that Jadzia and Dax differ in their beliefs: Jadzia believes that Dax has not engaged in criminal activity, but Dax believes that Dax has engaged in criminal activity. Thus, Dax and Jadzia hold contradictory beliefs at the same time, and these beliefs may enter into separate reasoning processes simultaneously. Jadzia and Dax have distinct 'belief boxes.' It is not simply that Jadzia has an explicit belief which Dax implicitly denies. They have distinct dispositional and explicit beliefs. We all sometimes explicitly aver one thing but act in some other way, as in the case of implicit racism. This, on its own, should not imply that we have two minds. If it did, the two-mind theory would be banal, since it would amount to the claim that humans are not perfectly rational or do not always act in accordance with their explicit beliefs. Thus, what is crucial to these simultaneous contradictory beliefs is that they are maintained *as the same kind of belief* (i.e. dispositionally, implicitly, explicitly, etc.). I will put this more formally below. Let belief$_k$ mean belief of some specific kind (i.e. dispositional, implicit, explicit, etc.).

> **Belief-K Condition:** If a subject believes$_k$ that p, then that subject does not believe$_k$ that not-p, unless that subject has two belief boxes.[3]

Steven Sloman, who endorses the two-system theory, while denying the two-mind theory (2014, 69; 1996, 3), has posited what he calls Criterion S, according to which, if a

[3] In order for it to be possible for 'dispositional beliefs' to be contradictory, one would have to distinguish sharply between dispositional beliefs and dispositions to believe (see Audi, 1994).

subject simultaneously believes contradictory propositions in responding to a reasoning problem, then there must be more than one reasoning system. Sloman takes this claim to be tautological. I will put this claim a bit more formally as follows:

> **Simultaneous Contradictory Belief (SCB) Condition:** A token reasoning process cannot decompose into two sub-processes operating simultaneously which result in *the generation of* simultaneous contradictory beliefs.

Elsewhere, some have argued that the SCB Condition is a way of distinguishing one-system accounts of human reasoning from parallel-competitive accounts of human reasoning (Osman 2004; Mugg 2013), but here I want to make a stronger point. The SCB Condition gives us a way to empirically distinguish one and two mind theories. Consider the following conditional:

> **Mind and Belief (MB):** If a thing has beliefs, then it is a minded thing.

If MB is true, then the SCB Constraint is as much about minds as it is about processes. Thus, if a cognitive system possesses beliefs, that cognitive system would constitute a mind.

If humans possess two minds in virtue of both the SCB Condition and MB being met, then both minds would be at the personal level. This is not trivial, since Frankish (2009) defends his two-mind theory by associating one mind with the personal level and the other with the sub-personal level. The issue of the relation between the personal/subpersonal distinction to the two mind theory is worth exploring in some detail here.

Frankish (2009) attempts to situate the System 1/System 2 distinction within the subpersonal/personal distinction. Briefly, a personal level/state/process/event is one that is ascribable to the person or creature as a whole (Dennett, 1987). A sub-personal level state/process/event is one that is not ascribed to the person or creature as a whole, but instead is ascribed to a part (or a subsystem) of that person or creature.[4] Frankish suggests that we identify S1 with sub-personal level attribution and S2 with personal level attribution. He gives us the following examples for personal and sub-personal reasoning. Suppose you are asked what is 21,582 divided by 11. If you are a math whiz, the answer may just come to you (1962). You would not, however, know how you

[4] 'Person' should be understood in a very minimal sense. Personal-level states are not sufficient for personhood, and do not themselves constitute the 'self.' Frankish is clear that he does not wish to imply otherwise (2009, 91).

worked out the answer. The process of determining the answer would be entirely sub-personal. However, most of us need to get out a pencil and paper and work through a series of steps. This process is personal, even though some steps along the way might be sub-personal (e.g. what is 22 divided by 11). The "defining feature" of personal reasoning is intentionality, by which Frankish merely means acting for reasons (2009, 92). Personal reasoning requires the use of working memory and is "therefore conscious" (2009, 93). However, the beliefs and desires motivating a particular instance of personal reasoning need not be conscious (i.e. they can be implicitly held).

Assuming that the sub-personal/personal distinction maps neatly onto the S1/S2 distinction, Frankish notes some important implications. First, S2 would not be a neural system in its own right, but is, rather, a virtual system "constituted by states and activities of the whole agent" (2009, 97). It is constructed out of sub-systems (2009, 99). He calls this an action-based view of S2 (2012, 42). Second, S2 is causally and instrumentally dependent on S1: instrumentally because S2 will use S1 subsystems to engage in autostimulation, whether it be inner speech, action simulation, or something else, and causally dependent because S1 (the sub-personal systems) generates the intentional actions used by personal reasoning. Lastly, S2 depends on S1 "to make its *outputs* effective" (2009, 97). That is, sub-personal "metacognitive attitudes make personal decisions effective" (2009, 98).

The difficulty for Frankish is that beliefs are personal level entities. *My brain* does not believe; *I* believe. *My reasoning system* does not reason; *I* reason. However, if two-mind theorists wish to use contradictory beliefs to argue for their position, then their claim would be that the two systems are the possessors of the contradictory beliefs. Minds that have reasoning systems have beliefs. So each mind has beliefs. Thus, once you endow certain cognitive systems with belief possession, they 'graduate' from being at the cognitive level to the agential level.

It is natural to interpret Jadzia and Dax as possessing beliefs at the personal level. Jadzia and Dax are two distinct systems *possessing* distinct beliefs, and as such we regard them as distinct minds at the personal level. In 'Dax', Odo and Sysco consider whether Curzon-Dax could have committed a murder he is accused of. Sysco, who knew Curzon-Dax for years, explains that Curzon could not have done it: he "knew the man." Odo replies "but did you know the symbiont inside the man?" Sysco and Odo characterize Jadzia, Curzon, and the Dax symbiont at the personal level. We are comfortable with distinguishing them at the personal level partly because they are distinct organisms, but my point here is that, if the SCB Condition or Belief-K Condition are supposed to aid in an argument for the two-mind theory, then the two-mind theorist must admit that the distinction between the two minds is not merely at the sub-personal (or cognitive) level. Instead, it is at the personal (or agential) level.[5]

So far, I have been offering sufficient conditions for Jadzia and Dax having two minds. I now turn to a necessary condition. Jadzia and Dax have some reduplication of parts. That is, Jadzia and Dax both have phenomenal states, propositional attitudes, and cognition. It is not as though Jadzia contains all the propositional attitudes and Dax possesses all the phenomenal states. If this were the case, then (plausibly) Jadzia and Dax would possess different *parts* of one mind rather than possessing distinct minds.

The two-mind theory operates at a higher level than dual-process theory or two-system theory. A mind can be a collection of systems. Evans explains:

> "[My] version of the two minds theory (Evans 2010b) makes the strong claim that there are two distinct forms of learning, memory and cognitive representation underlying the operations of the intuitive and reflective minds. There are implicit, procedural and habit learning systems in the old mind which can regulate our behavior without intervention by working memory, and which register no more than emotional or metacognitive feelings in consciousness" (Evans 2011, 91)

The idea is that there is a duplication of the various kinds of systems—humans possess two systems for learning, two for memory, two for mindreading, and (perhaps) even two for perceptual domains like vision. The two-mind theory is meant as a way to unify these dual-process and two-system accounts from various domains of psychology. The old mind has its own form or system of learning, memory, mindreading, and reasoning and the new mind has its own. If humans did have two minds, we should expect to find just such a duplication—just as in the Jadzia-Dax case. Thus, duplication of systems is a necessary condition on the two-mind theory.

[5] The forgoing discussion is not the case for two-system theorists wishing to make use of merely the SCB Condition. The two-system theorist denying the two-mind theory can say that it is misleading to say that, according to the dual-process theorist, beliefs are held at the Type-1 level, or to say that System 1 or System 2 *believe* anything. Supposing that there are two distinct processes, the picture, as they would have it, is that Type-1 and Type-2 processes (subpersonal and personal reasoning respectively) both issue a response, and these responses can be in contradiction with one another. However, both of these responses must be attributed to the *organism as a whole*, given that they are beliefs. That is, they are attributed at the *personal* level. Two-system theorists wishing to deny that the two-mind theory can simply reject the claim that the beliefs are stored separately.

Furthermore, there is good reason to think that duplication is a sufficient condition as well—that Jadzia-Dax possesses two reasoning systems, two perceptual systems, two mindreading systems, etc. seems to imply that Jadzia-Dax has two minds. However, we must be careful not to assume that the existence of a duality in one domain will correspond to the duality in another. It is crucial to the duplication in Jadzia-Dax's case that the duplicated systems cluster—one reasoning system is Jadzia's, the other is Dax's, one perceptual system is Jadzia's, the other is Dax's. That is, if the two-mind theory is true, then the two systems of various domains of psychology should not cross-cut one another. Furthermore, all the system's of Jadzia's interact with a much higher frequency than they interact with Dax's systems—Jadzia is one cognitive system, Dax is another. Call this the Duplication Principle:

> **Duplication Principle**: X has two minds, M_1 and M_2, if and only if there is a duplication of systems such that for each duplicated system S_1 and S_2, S_1 is a system of M_1 and S_2 is a system of M_2.

Thus, we have four ways that two-mind theorists could argue for their account. First, they might find evidence for a double dissociation between the two minds. Second, they could argue that the beliefs of the same kind are maintained simultaneously by single subjects (Belief-K Condition). Third, they could accept the SCB Constraint combined with MB and argue that simultaneous reasoning processes generate contradictory beliefs, which are maintained by separate systems. Finally, and most importantly, humans have two minds if and only if human cognitive faculties are duplicated. Having gotten clear on what it would take for there to be two minds, we may now turn to empirically motivated accounts of human reasoning.

3. One-System, Default-Interventionism, and Parallel-Competitive Theories

Here I will outline two versions of dual-process theory and contrast them with a one-system alternative, arguing that they should be understood as a continuum with borderline cases rather than admitting of sharp boundaries. I will begin with one-system accounts. There have been a number of models suggested. Human reasoning might be entirely rule-based, consisting of a complex structure of heuristics (see Kruglanski & Gigerenzer, 2011), or human reasoning might exist along a continuum, rather than as a bifurcation. Osman's (2004) one-system alternative is an extension of Cleeremans and Jimenez's (2002) dynamic graded continuum (DGC) theory of learning. On this connectionist account, implicit, automatic, and explicit processing form a continuum. Implicit reasoning, when they encounter novel reasoning problems, "involves making a set of abstractions or inferences without concomitant awareness of them" (995, see also 996). In contrast to implicit (but not automatic) reasoning, subjects have awareness in *explicit reasoning,* and this awareness "can be expressed as declarative knowledge" (995). Finally, automatic reasoning is "deliberately acquired through frequent and consistent activation of relevant information that becomes highly familiarized" (995). On her account, explicit processes may become automatic (in her sense) over time. However, automatic and explicit processes do not become implicit over time. That is, an explicatable process may become highly familiarized, but does not eventually occur outside awareness.

Dual-process accounts come in two varieties. First, according to default-interventionism, subjects default to one kind of processing and only sometimes use the second kind. Default-interventionism is the most common dual-process position (held by Kahneman, Frederick, Stanovich, and Evans). Second, on parallel-competitive accounts, the two processes operate at the same time and are in direct competition with one another. Because Type-1 processing is faster than Type-2 processing, it "always has its voice heard" (Sloman 1996, 3). The two processes are like racing horses, but the slow and steady Type-2 does not generally win the race.[6]

Parallel-competitive accounts might seem qualitatively distinct from default-interventionist accounts, since the two processes operate independently and at the same time on parallel-competitive models. There is indeed a position to be had here. However it is one that is 1) is an extreme version of parallel-competitive (and implausible given the empirical data), and 2) is a theoretical position that no one actually holds. Instead, parallel-competitive theorists think that the two processes causally interact in important ways. In fact, two parallel-competitive theorists, Frankish (2004; 2009; 2012) and Carruthers (2009; 2011), argue that System 2 is a virtual system that is realized in the cycles of System 1. That is, the processes that System 1 carries out are *constitutive* of the processes carried out by System 2.[7] On virtual system parallel-competitive accounts,

[6] Some have argued that parallel-competitive accounts cannot account for instances where Type-2 does win out, but this is a misunderstanding of the position. Parallel-competitive theorists can say that although Type-1 processing will end first, the subject may 'hold off' in responding until Type-2 processing has generated a response. Since Type-1 processing is automatic, as long as the stimulus is present, it will continue generating its response. Thus, when Type-2 processing completes the task, there is a fresh Type-1 response to compete with it.

[7] Frankish and Carruthers have an internal debate as to whether or not System 2 possesses its own mental states. Carruthers thinks that all the causal work is done by S1, and so S2 has no states of its own. Frankish disagrees, arguing that S2 has *sui generis* belief states.

System 2 is fully dependent on System 1. However, System 1 is not dependent on System 2. Now, the more the two processes interact, the closer the parallel-competitive model moves to a default-interventionist account. Thus, parallel-competitive and default-interventionist accounts are not sharply distinguished. Some accounts are border-line.

Now compare one-system accounts and default-interventionism. Again, the question is how integrated the two processes are. Suppose all processing initially is Type-1 processing, and only sometimes does Type-2 processing even come online, though when it does Type-1 processing shuts down completely. Perhaps this is different in a principled way from one-system accounts. However, suppose that Type-2 is dependent on Type-1 processing for its input (as Evans [2011, 94] and Stanovich [2011, 62] claim). Then it is less clear why we should regard these as distinct processes rather than parts of a more general process (see Kruglanski, 2013 for a similar point). Thus, we run into the infamous grain problem (Atkinson & Wheeler, 2003; 2004). As it applies here, the question is whether there is some level of description under which it is plausible (but not trivial) that there are reasoning processes that are distinct and not mere parts of a larger process. There are two related worries: first, how to determine whether two token processes are in fact sub-processes of a coarser-grained token process (call this the 'token grain-problem'); second, how to determine whether two types of processes are in fact sub-processes of a coarser-grained type of process (call this the 'type grain-problem'). If the grain-problem cannot be resolved, then default-interventionism and one-system accounts admit of vagueness.

The SCB Condition provides a principled way of distinguishing reasoning processes. One reason to think that the two processes are not parts of a more general *reasoning* process is that they can produce SCB. However, it is not clear that default-interventionism is compatible with the existence of SCB. Evans and Stanovich (2013) disagree with Sloman's "contention that simultaneous contradictory belief is a necessary condition for the existence of dual processes in conflict (his Criterion S)" (227). This disagreement should not be surprising, since default-interventionism does not conceive of the two processes in direct competition with one another. Rather, subjects default to Type-1 processing, which is sometimes overridden by Type-2 processing.[8] Thus, default-interventionists need some other way to solve the grain problem if they are to be distinguished from one-system accounts.

[8] If default-interventionism is not compatible with simultaneous contradictory belief, then the SCB Condition can empirically distinguish parallel-competitive accounts from both one-system and default-interventionist accounts

Default-interventionist accounts have moved increasingly toward one-system accounts, rather than more sharply distinguishing themselves from such one-system accounts. Recently Evans and Stanovich (2013) offered a revision of their accounts in response to criticisms. Evans's model has it that Type-1 reasoning automatically generates a response, but that then Type-2 reasoning reflects on this response, and (in conjunction with the amount of cognitive resources available and motivational factors) sets the amount of effort that the subject will use in assessing the response. As a result, all reasoning responses go through *both* kinds of processing. Why regard these as separate reasoning processes, rather than one reasoning process? I grant that there is nothing inconsistent with regarding them as distinct processes, but Evans gives us no reason to think that his revised account is still a dual-process account rather than a fleshed out one-system theory. In fact, Kruglanski (2013), in his commentary on Evans and Stanovich (2013), points out that Evans's account is remarkably similar to Gigerenzer's one-system account, according to which subjects have a toolbox of rule-based heuristics. Thus, there are accounts that are on a borderline between default-interventionism and one-system theory, such as Evans and Stanovich's (2013).

Are there borderline cases between parallel-competitive and one-system accounts? I think not. According to parallel-competitive accounts, not all reasoning results in Type-2 processing. Furthermore, the two processes are in competition with one another, which seems to give us a principled reason for distinguishing the processes. Finally, parallel-competitive theorists can use my SCB Constraint to distinguish their accounts from one-system accounts. Frankish and Sloman (but not Carruthers) accept the existence of SCB arising from the distinct processes (and, in their case, systems). Thus, there are principled ways of distinguishing parallel-competitive and one-system theories.

4. Two-Mind Theory cannot be Default-Interventionist

Philosophers and psychologists in the dual-process literature generally assume that both default-interventionism and parallel versions are compatible with two-mind theory. For example, Frankish, a parallel-competitive dual-process theorist, is a two-mind theorist, as are Evans and Stanovich,[9] both default-interventionist theorists. I will argue that default-interventionism is not compatible with the strong two-mind theory. If I am

[9] Stanovich is actually a 3-mind theorist. On his account there is the collection of module-like systems, (The Autonomous Set of Systems, or TASS), the algorithmic mind, and the reflective mind. TASS carries out only Type-1 processing, while the other two minds carry out both Type-1 and Type-2 processing. See Stanovich 2011, 62 for details concerning the relation between these minds.

right, then Stanovich and Evans must either become parallel-competitive theorists or reject the two-mind theory.

In section 2, I replied to an objection that Jadzia-Dax does not have two minds because the two are too integrated. I replied that the Jadzia and Dax's cognition is doubly dissociable, and neither is dependent on the other. As Selin Peers (a Trill expert) puts it, the process "is a joining. It is a total sharing, a blending…Neither is suppressed by the other" ('Dax'). However, as we examine default-interventionism, we find that the two systems/processes are too integrated to be partly constitutive of distinct minds.

In the previous section, I argued that the distinction between default-interventionism and one-system theory is vague. However, no vagueness arises between one-mind theory and two-mind theory, and the one-system theory is incompatible with the two-system theory (by the Duplication Principle). Therefore, default-interventionism is not compatible with two-mind theory.

Here is another way to see the objection. Sloman rightly claims that for something to be a system, "a set of cognitive processes and representations must have some individual autonomy; they must operate and compute independently enough that they can be held responsible for critical aspects of behavior" (2014, 71). If systems must have some individual autonomy and operate fairly independently, then the same can be said for minds. A mind is, after all, a kind of system. However, on Evans's account, all reasoning goes through both Type-1 and Type-2 processing. Type-1 generates a response, then Type-2 determines whether to simply accept that response or undergo further Type-2 processing that would potentially override the Type-1 response (see Evans 2011, 94). Thus, neither process can, on its own, be responsible for some critical aspects of behavior.

Default-interventionism is incompatible with Belief-K Condition being met or the SCB Condition combined with MB being met. First, note that default-interventionism has it that Type-1 and Type-2 responses are generated at different times: first the Type-1, then (sometimes) the Type-2. Thus, they cannot use the SCB Condition, since the SCB Condition require that the beliefs are *generated* simultaneously. Second, default-interventionism has it that Type-2 processing (sometimes) *overrides* or *intervenes* on Type-1 responses, rather than generating responses all on its own in addition to Type-1 responses. If the intervention is successful, then the Type-2 response replaces the Type-1 responses. Thus, subjects will not have contradictory beliefs of the same kind at the same time: the Belief-K Condition will not be met if default-interventionism is true. Of course, Belief-K Condition, SCB Condition, and MB are only sufficient conditions for two-system theory. So it does not follow from my argument here that default-interventionism is incompatible with the two-mind theory. However, it does imply that default-interventionists will have to argue for the two-mind theory in some other way.

There is another, deeper problem for the combination of default-interventionism and the two-mind theory: default-interventionism cannot satisfy the Duplication Principle. Above I said that it is not as though Jadzia and Dax divide their labor: Jadzia doing all the perceptual work and Dax doing all the cognitive work, say. Instead, there is a duplication of system types. Remember, that two-system and dual-process accounts exist in the various domains does not, on its own, imply that they are all gesturing at the same two minds. It may be that the two-system and dual-process accounts in various domains of psychology are merely *employing a similar strategy* for explaining complex data rather than pointing to different parts of the *same* two minds. If humans possess two minds, then the dual-process and two-system accounts from diverse domains of psychology should fit well together. This is what we would expect if we could empirically investigate Jadzia-Dax: Jadzia's cognitive systems would align, and so would Dax's. Thus, if the two-mind theory is true, we should find organizational and structural similarities between the dual-process and two-system theories in each domain.

Let us turn to the empirical literature. Our question is to what extent the old/new mind distinction cross-cuts the various two-system and dual-process accounts of the domains of psychology. Unfortunate for Evans, it seems that there is a fair bit of cross-cutting. The kind of cross-cutting I have in mind here differs from the cross-cutting offered against dual-process theories of reasoning, according to which the *properties* used to distinguish Type-1 and Type-2 reasoning cross-cut each other (see Carruthers, 2013; J. S. B. Evans, 2008; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Mugg, forthcoming). Here the claim is that the various system 1/system 2 or Type-1/Type-2 processes of theories from different domains cross-cut one another. Thus, there is good reason to think that the dual-process and two-system accounts across the subfields of psychology are not pointing to the same two minds. From the above section, we already have a good handle on dual-process accounts within reasoning. In the remainder of this section, I will outline dual-process theory within social cognition and mindreading and argue that the duality in these areas do not correspond to alleged Type-1and Type-2 processing in reasoning.

4.1 Social cognition (Smith and DeCoster)

Social psychologists have proposed many dual-process models to explain *specific* tasks, and some have gone further in attempting to develop a *general* dual-process account of social cognition that accommodates these specific dual-process models. I will focus on Smith and DeCoster's (2000) influential account, which Evans and Stanovich both cite frequently. Smith and DeCoster (2000) draw heavily on the associative/rule-based distinction, arguing that associations and rules are "two separate memory systems" (Smith & Collins, 2009; 200). They write, "in brief, what we term the *associative processing mode* is based directly on the properties of the slow-learning system…in contrast, the *rule-based processing mode* uses symbolically represented and culturally transmitted knowledge as its 'program'" (110). Smith and DeCoster aim to unify several dual-process theories from social cognition. Their account seems to be parallel-competitive, rather than default-interventionist, since they "assume that the two processing modes generally operate simultaneously rather than as alternatives or in sequence" (Smith and DeCoster, 112). However, they also "do not see that distinction [between default-interventionist and parallel-competitive accounts] as very clear-cut" (Smith & Collins 2009; 205). They emphasize conscious control and effortfulness as a common theme in rule-based processing across dual-process theories of social cognition (125).

Problematically, Evans and Stanovich are clear that the associative/rule-based distinction must be discarded, as they concede to Kruglanski and Gigerenzer (2011) that putative associative processes can always be captured by rules. As Evans (2006) puts it, "I am not sure it is wise to describe System 2 as 'rule-based'…if only because it implies that System 1 cognition does not involve rules" (204, quoted in Evans and Stanovich (2013), 231). Smith and DeCoster (2000) are aware that

> "associations sometimes have been termed 'rule', [but] for clarity it is important to preserve the distinction between associations (which are built up through repeated experiences over time and are not necessarily interpersonally shared or symbolically encoded) and rules (which can be explicitly learned on a single occurrence and are symbolically represented and often interpersonally shared)" (111)

Thus, Smith and DeCoster conceive of associations and rules as *qualitatively* distinct kinds of processing, and this forms the basis of their conciliatory dual-process account of social cognition. Since Evans and Stanovich reject a characterization of Type-1 and Type-2 processing using the associative/rule-based distinction, it is unlikely that these theories are gesturing toward the same two minds.

There are other problems for grouping Smith and DeCoster's (2000) account with Evans and Stanovich's accounts. According to Smith and Collins (2009), rule-based processing can "*effortlessly override* the automatic activation of stereotypes by accessing and considering their relatively more favorable 'personal beliefs' about the group's characteristics" (199, emphasis mine). This is in direct contradiction to Evans and Stanovich, who claim that Type-2 processing is *necessarily effortful* and that Type-1 is *necessarily not effortful*. Furthermore, according to Smith and DeCoster, when rule-based processing occurs, it "generally gives rise to a higher level of perceived validity of the conclusion or judgment and to more long-lasting effects" (201). Although they deny the 'quick and dirty' characterization of associative processing (since biases can result from "motives, by priming, or by other factors (e.g. current mood" (206)) it is not the case that using rule-based processing will result in less biases. In fact, they claim that "research in social psychology demonstrates that intentional efforts to correct bias may even lead to further bias" (207) (see Wegener & Petty 1997). This is in sharp contrast to most default-interventionists, who claim that the feeling of rightness is generated by Type-1 reasoning (Thompson, 2009, 176). Kahneman goes so far as to say that after he adopted a new policy of marking in order to avoid the anchoring effect when grading tests he was "less happy with and less confident in [his] grades…but…recognized that this was a good sign, an indication that the new procedure was superior" (2011, 84). To put it in default-interventionist, theory-laden terms, lacking a feeling of rightness implies that the result is not generated by Type-1 processing, and as such, must have come from Type-2 processing, which generates no feeling of rightness.

As I have argued above, key to seeing whether an account is really dual-process or single system is the interaction between the putative two systems or processes. Smith and DeCoster's account, at times, look remarkably similar to Osman's (2004) one system account, which is supposed to be a rival to dual-process theory. Smith and DeCoster (2000) are clear that repeated symbolic rule use can "create the conditions for associative learning…With enough practice, therefore, the answer to such a problem just pops into consciousness" (115-116). These points are repeated in Smith and Collins (2009), who write that "repeated use of symbolic rules creates the conditions for associative learning, so eventually the same answer that is generated by the rule-based system can be retrieved by pattern-completion in the associative system" (202). Conversely, associative information can become symbolic rules: "people can reflect on their own past experiences and summarize them, perhaps in the form of a symbolically represented rule" (Smith and DeCoster 2000, 116).[10]

[10] Smith and DeCoster (2000) seem to admit the existence of simultaneous contradictory belief, following Sloman (1996). The existence of SCB would be incompatible with Osman's (2004) account, as she herself says (though she and others have argued that Sloman has failed to support the existence of SCB (see also

Recall that, on Osman's one-system account, explicit processes may become automatic (though not implicit) over time. Smith and DeCoster's (2000) account seems similar in that rule-based processing may become associative processing just as explicit processing may become automatic. Such transitions would be less likely on Evans and Stanovich's accounts, where the Type-1 processing is carried out by module-like systems.

I admit that Smith and DeCoster's account might fit well with some dual-process theories in cognitive psychology (particularly parallel-competitive accounts like Sloman's (1996). See Smith and DeCoster (2000), 123). However, it is not the case that Smith and DeCoster's account fits well with Evans and Stanovich's account. To be sure, Smith and DeCoster's account bears some resemblance to Evans and Stanovich's accounts, but, then again, Smith and DeCoster's account also bears some resemblance to Osman's account as well. Thus, it is unlikely that Smith and DeCoster's associative and rule-based modes correspond to the default-interventionist's two minds.

4.2 Mindreading (Apperly and Butterfill)

Apperly and Butterfill have developed a two-system account of mindreading—the ability to attribute mental states to others. The difficulty philosophers and psychologists face when interpreting the empirical data is that there is evidence that nine-month old humans attribute mental states (based on looking-time paradigms. See, e.g., Onishi & Baillargeon, 2005), but children are unable to pass false-belief tasks (such as the Sally-Anne Test) until three or four years (Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). Responses to this seemingly contradictory evidence in the literature tend "to be polarized: Infants [and] nonhuman animals…either employ mental concepts such as perception or belief or get by exclusively with behavioral rules" (I. A. Apperly & Butterfill, 2009, 966). This polarization "might be resolved in one of two ways: either one set of evidence would prove to be unsound, or the apparent contradictions actually reflect genuine diversity in" human mindreading (Apperly, 2011; 133). Apperly clearly takes himself to be employing a *strategy* for resolving complex data similar to dual-process theorists in other domains, rather than gesturing at a two-mind architecture:

> "These dual requirements [for solving the problem of unbounded information processes (i.e. the frame problem)] are not unique to mindreading, and for topics as diverse as social cognition (e.g. Gilbert, 1998), number cognition (e.g., Feigenson, Dhane & Spelke, 2004), and

Mugg 2013).

> general reasoning (e.g., Evans, 2003), there is strong evidence that both kinds of solutions are employed. I suggest that the same is true for mindreading (see Apperly and Butterfill, 2009)" (133)

Thus, introducing two systems is not intended as gesturing to the *same two systems* (or minds) found in other areas of psychology. Instead Apperly and Butterfill are employing a strategy found in other areas of psychology for explaining how a processing in a domain can "be both flexible and efficient" (Apperly and Butterfill 2009, 957). Apperly and Butterfill write that they "advocate a view based on *lessons* from another domain" (2009, 953, emphasis mine). However, they make no effort to say how their account fits with these other theories, and admit that the details of the various theories differ. How much do they differ? I will argue that Apperly and Butterfill's account differs from Stanovich and Evans's two mind account in ways that indicate that they are not implicating *the same two minds or systems*.[11]

On Apperly's (2011) account, there is a low and high level of mindreading. The lower level is present in infants and non-humans and does not involve language. Low level mindreading is fast, and it uses a distinct set of concepts, which can *track* goals, beliefs, and desires without representing them as goals, beliefs, and desires, *as such*. High level mindreading is the full-blown mindreading measured by false-belief tasks like Sally-Anne. It tends to be language involving (but does not appear to be "critically dependent on the availability of grammatically structured language" (159)), is more flexible, and is slower. Apperly claims that high and low-level mindreading are "at least partially dissociable" (167), given evidence from autistic subjects, since their "high-level mindreading abilities might not be atypical" (167).

Some of the properties Apperly uses are certainly familiar from dual-process theories of reasoning—the fast/slow and the evolutionarily old/new distinctions in particular. However, there are some major differences. First, notice that low level mindreading is supposed to lack language. However, in dual-process theories of reasoning, both Type-1 and Type-2 processing are language involving. Otherwise Type-1 processing would not be implicated at all in cases like the conjunction fallacy or belief-bias. Second, high and low level mindreading are supposed to be "at least partially dissociable" (167). However, on default-interventionism, Type-2 processing is dependent upon Type-1 processing for its input. Thus, they will be dissociable only in that Type-1 processing can

[11] This is not meant as an objection to any of these theories as such, or to the compatibility of these theories. My claim is merely that Evans and Stanovich's Type-1/Type2 distinction and Apperly and Butterfill's low/high level mindreading are not governed by the same systems or minds.

occur without Type-2 processing. Type-2 processing cannot occur without at least some Type-1 processing. Thus, the relation of the two kinds of mindreading and two kinds of reasoning differ.

A point of commonality between Evans and Stanovich's accounts and Apperly and Butterfill's accounts is the importance of working memory or executive functioning. Recall that Evans and Stanovich (2013) claim that the distinction between Type-1 and Type-2 processing is the distinction between a process being autonomous or working-memory involving. Consider cognitive decoupling, which Stanovich and Evans both emphasize in their accounts: subjects make a copy of a representation, which is kept separate from one's beliefs (i.e. in working memory) such that it can be manipulated. This task of keeping the two separate is cognitively taxing (Leslie 1987). That is, it takes executive functioning. Similarly, executive function plays a central role in Apperly's account. He writes:

> "unlike the cases of language there is equally clear evidence that executive function continues to have a significant role in the mindreading abilities of adults…However, there is also good evidence that some mindreading processes are much less effortful and resource demanding…and there is evidence that adults can implicitly and automatically calculate what someone else sees (Level-1 visual perspective-taking)" (111)

Working memory and executive functioning are distinct, but closely related. Working memory is that which temporarily stores and manipulates information. For example, subjects use their working memory to remember the pattern in a dot matrix (De Neys, 2006). Executive functioning, in contrast, is that which inhibits or suppresses action tendencies or mental states. For example, bilinguals of audible languages use their executive function when they speak—they must suppress their non-active language (Bialystok & Viswanathan, 2009; Moreno et al., 2011).

One might think that the close relationship between working memory and executive functioning and the important role that it plays in these theories is evidence that dual-process accounts in these two domains are converging. However, the automatic/working-memory distinction alone is not sufficient for showing a robust convergence. The difficulty is that if the two-mind theory amounts to the claim that some processes involve working memory whereas others do not, then the two-mind theory is banal. Any one-system theorist would agree that not all cognitive processes involve working memory or executive functioning (see Mugg, forthcoming). Thus, even if we ignore all the other ways in which dual-process theories in reasoning and mindreading do not converge, working-memory/executive function involving is little evidence for a convergence.

4.3 Conclusion

If the two-mind theory is true, then there should be a duplication of systems—two mindreading systems, two reasoning systems, two perceptual systems, etc. However, default-interventionism, at least of Evans and Stanovich's kind, relates the processes too tightly to suggest a clear bifurcation of systems. Furthermore, we do not find a deep commonality between the dual-process theories across domains of psychology. Evans himself seems aware of the problem of mapping his own dual-process account onto his own two mind account. He admits that "there are Type-1 processes operating within both the old and new minds" (2011, 93). So, as it turns out, even his own default-interventionist account does not perfectly line up with his two-mind theory.[12]

5. Conclusion

I have argued that the typical way of understanding the relation between the various dual-process and two-mind theories is mistaken in two ways. First, although I agree that dual-process theories divide into parallel-competitive and default-interventionist versions, I have argued that these two lie on a continuum with one-system accounts (especially the one-system dynamic-graded continuum account). Second, given the conditions gleaned from the two-minds of Jadzia-Dax, it is clear that default-interventionism is incompatible with the two-mind theory for three reasons. First, default-interventionists deny the possibility of simultaneous contradictory belief (as many one-system theorists do). Thus, they cannot use the combination of the SCB Condition, Belief-K Condition, and MB Condition to support their claim that humans have two minds. Second, default-interventionism cannot meet the Duplication Principle—which is necessary and sufficient for the two-mind theory. Third, default-interventionist accounts of human reasoning (such as Evans and Stanovich's) do not fit with dual-process theories in other domains of psychology (such as mindreading and social cognition). It is implausible that the theories

[12] A two-mind theorist might reply that the various dual-process theories do have *some* resemblance. Namely, they all draw some properties from the 'standard menu.' However, Evans and Stanovich have recently abandoned the standard menu as a way of distinguishing the two processes because of the existence of cross-cutting. The standard menu did serve to unify the various dual-process accounts, but that recourse is not available to Evans and Stanovich (see Mugg, forthcoming).

from these diverse areas of psychology are gesturing at the same two minds. More likely, they are merely employing a similar explanatory strategy.

References

Apperly, Ian. 2011. *Mindreaders: The Cognitive Basis of "Theory of Mind."* New York: Psychology Press.

Apperly, Ian. A., and Stephen. A. Butterfill. 2009. "Do Humans have Two Systems to Track Beliefs and Belief-like States?" *Psychological Review* 116 (4): 953–970.

Atkinson, Anthony. P., and Mark Wheeler. 2003. "Evolutionary Psychology's Grain Problem and the Cognitive Neuroscience of Reasoning." In *Evolution and the Psychology of Thinking: The Debate*, edited by Daniel. E. Over, 61–99. New York: Psychology Press.

Atkinson, Anthony. P., Mark Wheeler. 2004. "The Grain of Domains: The Evolutionary-Psychological Case Against Domain-General Cognition." *Mind and Language* 19 (2): 147–76.

Audi, Robert. N. 1994. "Dispositional Beliefs and Dispositions to Believe." *Noûs* 28 (4): 419–34.

Bialystok, Ellen., and Mythili Viswanathan. 2009. "Components of Executive Control with Advantages for Bilingual Children in Two Cultures." *Cognition* 112 (3): 494–500.

Carruthers, Peter. 2009. "Architecture for Dual Reasoning." In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 109–27. Oxford: Oxford University Press.

Carruthers, Peter. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Carruthers, Peter. 2013. "The Fragmentation of Reasoning." In *La Coevolución de Mente y Lenguaje: Ontogénesis y Filogénesis*, edited by Quintanilla. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

De Neys, Wim. 2006. "Dual Processing in Reasoning: Two Systems but One Reasoner." *Psychological Science* 17 (5): 428–433.

De Neys, Wim. 2012. "Bias and Conflict: A Case for Logical Intuitions." *Perspectives on Psychological Science* 7 (1): 28–38.

Dennett, Daniel. C. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Evans, Jonathan. St. B. T. 2006. "The Heuristic-Analytic Theory of Reasoning: Extension and Evaluation. *Psychonomic Bulletin & Review* 13 (3): 378–395.

Evans, Jonathan. St. B. T. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59: 255–278.

Evans, Jonathan. St. B. T. 2009. "How Many Dual-Process Theories do We Need? One, Two, or Many?" In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 33-54. Oxford: Oxford University Press.

Evans, Jonathan. St. B. T. 2010. *Thinking Twice: Two Minds in One Brain*. New York: Oxford University Press.

Evans, Jonathan. St. B. T. 2011. "Dual-Process Theories of Reasoning: Contemporary Issues and Developmental Applications." *Developmental Review* 31 (2): 86–102.

Evans, Jonathan. St. B. T., and Keith Stanovich. 2013. "Dual-Process Theories of Higher Cognition Advancing the Debate." *Perspectives on Psychological Science* 8 (3): 223–241.

Fiala, Brian, Adam Arico, and Shaun Nichols. 2011. "On the Psychological Origins of Dualism: Dual-process Cognition and the Explanatory Gap." In *Creating Consilience: Issues and Case Studies in the Integration of the Sciences and Humanities*, edited by Edward Slingerland and Mark Collard, 88–110. Oxford: Oxford University Press.

Frankish, Keith. 2004. *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, Keith. 2009. "Systems and Levels: Dual-System Theories and the Personal-Subpersonal Distinction." In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 89–108. Oxford: Oxford University Press.

Frankish, Keith. 2010. "Dual-Process and Dual-System Theories of Reasoning." *Philosophy Compass* 5 (10): 914–926.

Frankish, Keith. 2012. "Dual Systems and Dual Attitudes." *Mind & Society* 11 (1): 41–51.

Frankish, Keith, and Jonathan St. B. T. Evans. 2009. "The Duality of Mind: An Historical Perspective." In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 1–32. Oxford: Oxford University Press.

Kahneman, Daniel. 2011. *Thinking: Fast and Slow*. New York: Macmillan.

Kahneman, Daniel, and Shane Frederick. 2002. "Representativeness Revisited: Attribute Substitution in Intuitive Judgment." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 49–81. Cambridge: Cambridge University Press.

Keren, Gideon, and Yaacov Schul. 2009. "Two Is Not Always Better Than One A Critical Evaluation of Two-System Theories." *Perspectives on Psychological Science* 4 (6): 533–550.

Kruglanski, Arie W. 2013. "Only One? The Default Interventionist Perspective as a Unimodel—Commentary on Evans & Stanovich (2013)." *Perspectives on Psychological Science* 8 (3): 242–247.

Kruglanski, Arie W., and Gerd Gigerenzer. 2011. "Intuitive and Deliberate Judgments are Based on Common Principles." *Psychological Review* 118 (1): 97–109.

Mallon, Ron, and Shaun Nichols. 2011. "Dual Processes and Moral Rules." *Emotion Review* 3 (3): 284–285.

Moreno, Sylvain, Ellen Bialystok, Raluca Barac, E. Glenn Schellenberg, Nicholas J. Cepeda, and Tom Chau. 2011. "Short-Term Music Training Enhances Verbal Intelligence and Executive Function." *Psychological Science* 22 (11): 1425–1433.

Mugg, Joshua. Forthcoming. "The Dual-Process Turn: Why Recent Defenses of Dual-Process Theory of Reasoning Fail." *Philosophical Psychology*. DOI: 10.1080/09515089.2015.1078458.

Mugg, Joshua. 2013. "Simultaneous Contradictory Belief and the Two-System Hypothesis." *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*: 1044-1048.

Nagel, Jennifer. 2011. "The Psychological Basis of the Harman-Vogel Paradox." *Philosophers' Imprint* 11 (5): 1–28.

Onishi, Kristine. H., and Renée Baillargeon. 2005. "Do 15-Month-Old Infants Understand False Beliefs?" *Science* 308 (5719): 255–258.

Osman, Magda. 2004. "An Evaluation of Dual-Process Theories of Reasoning." *Psychonomic Bulletin & Review* 11 (6): 988–1010.

Sloman, Steven. A. 1996. "The Empirical Case for Two Systems of Reasoning." *Psychological Bulletin* 119 (1): 3–32.

Sloman, Steven. A. 2014. "Two Systems of Reasoning, an Update." In *Dual Process Theories of the Social Mind*, edited by Jeffery Sherman and Bertram Gawronski, 69–79. New York: Guilford Press.

Smith, Eliot. R., and Elizabeth C. Collins. 2009. "Dual-Process Models: A Social Psychological Perspective." In *In Two Minds: Dual Processes and Beyond*, edited by

Jonathan St. B. T. Evans and Keith Frankish, 197–216. Oxford: Oxford University Press.

Smith, Eliot. R., and Jamie DeCoster. 2000. "Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems." *Personality and Social Psychology Review* 4 (2): 108–131.

Stanovich, Keith. 2011. *Rationality and the Reflective Mind*. New York: Oxford University Press.

Thompson, Valerie. 2009. "Dual Process Theories: A Metacognitive Perspective." In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan St. B. T. Evans and Keith Frankish, 171–196. Oxford: Oxford University Press.

Wegener, Duane. T., and Richard Petty. 1997. "The Flexible Correction Model: The Role of Naive Theories of Bias in Bias Correction." *Advances in Experimental Social Psychology* 29: 142–208.

Wellman, Henry M., David Cross, and Julanne Watson. 2001. "Meta-analysis of Theory-of-mind Development: The Truth about False Belief." *Child Development* 72 (3): 655–684.

Wimmer, Heinz, and Josef Perner. 1983. "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." *Cognition* 13 (1): 103–128.

Journal of Cognition and Neuroethics

"Your Body Has Made a Different Choice": Cognition, Coercion, and the Ethics of Consent in Octavia E. Butler's *Lilith's Brood* and *Fledgling*

Meghan K. Riley
University of Waterloo

Biography

Meghan K. Riley is a PhD student in English Language and Literature at the University of Waterloo in Waterloo, Ontario. Her research interests include speculative fiction, critical race pedagogy, intersectionality, feminist standpoint theory, and the ethics of care.

Acknowledgements

I would like to thank Victoria Lamont for her critique of the earliest version of this paper. I would also like to thank the participants of the March 2015 conference "The Work of Cognition and Neuroethics" for their comments on that early draft, as well as two anonymous reviewers for their substantial and helpful suggestions.

Publication Details

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

Citation

Riley, Meghan K. 2015. "'Your Body Has Made a Different Choice': Cognition, Coercion, and the Ethics of Consent in Octavia E. Butler's *Lilith's Brood* and *Fledgling*." *Journal of Cognition and Neuroethics* 3 (3): 113–137.

“Your Body Has Made a Different Choice”: Cognition, Coercion, and the Ethics of Consent in Octavia E. Butler’s *Lilith’s Brood* and *Fledgling*

Meghan K. Riley

Abstract

Power is a common theme in Octavia E. Butler’s novels and short stories. The majority of the unequal power relationships are initiated and sustained through sex, sexual attraction, biochemical addiction, and mind control via biochemical influence and/or pheromones. The emphasis on coercion and subterfuge, as well as the association between sex and brain chemistry, allows for a critical consideration of Butler’s work as bearing upon debates over rape, medical ethics, and consent. Situated within a framework that includes a discussion of Kant’s the formula of the end in itself as well as of informed consent in bioethics, this article attempts to address features of Butler’s work which have gone largely unexamined within a philosophical context.

Keywords

Ethics, Coercion, Consent, Rape, Cognition, Kant, Bioethics, Science Fiction, Octavia E. Butler

> All struggles are essentially power struggles. Who will rule? Who will lead? Who will define, refine, confine, design? Who will dominate?
> —Octavia E. Butler, Parable of the Sower[1]

> I began writing about power because I had so little.
> —Octavia E. Butler, Interview with Carolyn S. Davidson, “The Science Fiction of Octavia Butler”

1. Introduction

Emphases on power are prevalent in Octavia E. Butler’s work. Butler is concerned, even obsessed, with issues of control, coercion, and consent. *Wild Seed’s* Doro demands that Anyanwu produce children with him, lest he harm the children she already has. In “The Evening and the Morning and the Night,” Lynn, who lives with Duryea-Gode disease, realizes that her particular pheromones allow her to draw men with the disease to her and influence them to follow her commands. The protagonist of *Kindred* travels

[1] This quote appears as the introductory parable of Chapter 9 in Butler’s *Parable of the Sower.*

back in time, only to find out that her great-great-grandmother was the product of rape. Anyanwu's granddaughter, Mary, can – somewhat like Lynn – attract and compel her relatives to do her bidding. The colonized humans of "Bloodchild" incubate alien children in return for an elixir which makes them young again, but are also never quite free of the aliens' seductive power; nor are the humans free from the power dynamics associated with being colonized.

Various treatises on Butler address power struggles across time and space in Butler's fiction. Sandra Govan addresses Doro's coercion of Anyanwu in *Wild Seed,* and in particular Doro's use of the "time-encrusted masculine ploy" to get Anyanwu pregnant, "the most immediate method he can use to control" her (1986, 85-86). Theri Pickens observes that "Butler's oeuvre stresses the impact of hierarchical relationships" (2014, 33), and Erin M. Pryor Ackerman notes that "[t]he issues of power and agency in Butler's writings have produced a wealth of criticism," (2008, 35). Some of the scholarship on Butler explicitly addresses power and desire in *Lilith's Brood* and *Fledgling*. Frances Bonner, for instance, links the relationship between desire, power, and consent in *Lilith's Brood* (1990, 58) as well as discusses both rape and forced reproduction. Marty Fink posits that "[a]s in 'Bloodchild' and *Dawn* … [in *Fledgling*] physical need and erotic transcendence preclude the possibility for escape," asking if "consent might not be plausible because of the factors informing [their] decisions" (2010, 418). Florian Bast suggests that "the possibility of agency is called into question when [Butler's characters are] confronted with biological realities rather than social constructions" (2010).

In this paper, I fully articulate the strategies that Butler's powerful characters utilize to control the less powerful, as they extend beyond "biological realities," desire, and sex. I show that far from relying exclusively on sexual coercion and drug addiction, Butler's powerful characters rely on a number of very human strategies to establish the unequal power dynamics. I also aim to make Butler's associations between sexual coercion, the ethics of consent, and medical ethics explicit, thereby further exposing the relevance of Butler's work for discussions on rape culture, women's reproductive rights, and bioethics. The following paper contextualizes Butler's treatises on the ethics of consent through a philosophical grounding. I first situate readers by providing a brief summary of Butler's work on the ethics of consent – namely, *Lilith's Brood* (2000) and *Fledgling* (2005) in section 1.2. Next I discuss how Butler explicitly and implicitly considers the ethics of consent in *Lilith's Brood* and *Fledgling* (sections 2 and 3), situate Butler's work in a larger discourse of consent and cognition (section 4), and explain the relevance of Butler's work to a discussion of rape culture, women's reproductive rights, and bioethics (section 5). In section 6, I conclude with a brief consideration of the relevance of the ethics of consent in

Butler's work, both in the body of existing scholarship on Butler, and for public discourse on rape, women's reproductive rights, and bioethics.

1.2 *Lilith's Brood* and *Fledgling*

Octavia E. Butler's two works to deal most prominently with the ethics of consent are *Lilith's Brood* and *Fledgling*. *Lilith's Brood,* originally published in 1989 as *Xenogenesis,* consists of *Dawn* (1987), *Adulthood Rites* (1988), and *Imago* (1989). The trilogy commences with Lilith's realization that she has been held captive by the alien Oankali for over two hundred years. The Oankali have periodically woken Lilith from suspended animation in order to determine her fit for her new roles – teaching the other humans how to survive on the post-apocalyptic Earth, and bearing the first Oankali-human hybrid, or construct. Once the Oankali decide that Lilith is indeed the human that they want, they begin to prepare her. What follows is a Cold War of sorts between Oankali and human resisters – humans who have been altered by the Oankali so they cannot have children on their own, but refuse to breed with the Oankali. Over the next thirty years, Lilith's children – human and Oankali hybrids, or constructs - must convince the Oankali to let human resisters have a separate colony and human children. Lilith's children must also find mates among the few willing humans left unclaimed by the older Oankali.

In *Fledgling,* Shori Matthews fights for her life against the old and influential white vampires who comprise the entrenched power system of the hidden vampire society and who feel threatened by a black vampire who can walk in the sun. *Fledgling,* as with *Lilith's Brood,* begins *in media res*. Its protagonist, like Lilith, awakes confused and alone, and questioning her sanity. The similarities end between Lilith and *Fledgling's* protagonist end there, however. Shori has awakened with amnesia, but soon learns that she is a 53-year-old Ina (Butler 2005, 70-72), or vampire. Shori appears to be a young black child, but she is much more like the Oankali than like Lilith. She holds most of the power in her relationship with her symbionts, the humans that she utilizes for sustenance and sex. It is Shori who withholds information, seduces, and coerces. As Shori attempts to find out who killed her family and left her with amnesia, her symbionts become pawns in a battle between the Ina factions.

Once the alien Oankali and the vampire Ina decide they want something, they take it. In this case, what they want are humans, who provide sexual release for both Oankali and Ina. Humans are also breeding partners in the case of the Oankali, and sustenance and servants for the Ina. The Oankali and Ina are not opposed to using force to subdue their "trade" partners (Butler 2000, 289) or "symbionts" (Butler 2005, 69), the Oankali and Ina's respective terms for the humans that they use. They are also not opposed to literally altering the humans' brain chemistry; indeed, it is their primary way of keeping the humans invested in the relationship. The Oankali inject humans with a biochemical substance which has the ability to calm humans (Butler 2000, 191 & 619) as well as encourage a human's body to "secrete specific endorphins" (Butler 2000, 512). The bite of an Ina serves the same function (Butler 2005, 79). Moreover, repeated exposure to an Ina's saliva means death if the human no longer has access to the saliva for a prolonged period of time (Butler 2005, 79-80). Human symbionts are addicted to, and dependent upon, the biochemical in the saliva (Butler 2005, 76 & 79).

2. Cognition, Coercion, Force, and Consent in *Lilith's Brood*

Though the Oankali's release of a human is not a death sentence, the Oankali pose many other dangers. The Oankali control humans through one of five strategies: they "read" human body language and scents, use physical force, present the humans with dichotomous "choices" in order to ensure an outcome favorable for the Oankali, omit information, and drug the humans. The first strategy, that of utilizing their extrasensory abilities to know human fears and desires, is one that the Oankali use often.

2.1 "They Know Our Bodies Better than We Do": "Reading" Humans in *Lilith's Brood*

The Oankali are particularly threatening because they perceive all that humans do not perceive about themselves. The Oankali are so perceptive that some humans think that the Oankali can read minds (Butler 2000, 25). The Oankali can even tell when humans are lying; because of their incredible senses, they "'can't help knowing'" when a human lies (Butler 2000, 619). The Oankali also perceive the humans' sexual attraction to and biochemical need for the Oankali:

> "You said I could choose. I've made my choice!"
>
> "You have, yes." It opened its jacket with its many-fingered true hands and stripped the garment from him. When he would have backed away, it held him. It managed to lie down on the bed without seeming to force him down. "You see. Your body has made a different choice." (Butler 2000, 189)

The Oankali's incredible extrasensory perception is ultimately only a foil for force. The Oankali routinely resort to physical force, which they invariably justify by saying that they only did what the humans wanted them to do.

2.2 "When he would have backed away, it held him": Forced Sexual Encounters and Reproduction between the Oankali and Humans

Physical force is, with few exceptions, a precursor to seduction for the Oankali. The Oankali assert that the humans want to be intimate with the Oankali, though their actions would seem to say otherwise. Lilith says, "They know our bodies better than we do" (Butler 2000, 169). The Oankali's defense of their sexual coercion of the humans is eerily similar to the arguments of rape apologists. Indeed, some of the men in the novel feel as though they have been raped by men or at least that they have been raped *like* women when they have intercourse with the third sex ooloi (Butler 2000, 192 & 203). As Rachel Pollack notes, rape is central, and apparently acceptable, in *Dawn,* the first book of *Lilith's Brood* (Pollack qtd. in Bonner 1990). Meanwhile, human women bear the brunt of the Oankali's efforts to transform the human species, with the Oankali again relying on their ability to "read" humans. Lilith tells Tino, a human resister who becomes her mate, her discomfort at failing to give her words meaning and impact:

> "They forced you to have kids?" the man asked.
>
> "One of them surprised me," she said. "It made me pregnant, then told me about it. Said it was giving me what I wanted but would never come out and ask for."
>
> "Was it?"
>
> "Yes." She shook her head from side to side. "Oh, yes. But if I had the strength not to ask, it should have had the strength to let me alone." (Butler 2000, 274)

It is Nikanj who impregnates Lilith and produces the first Oankali-human construct. Nikanj makes Lilith pregnant without her verbal consent, or even her knowledge, and it uses force to make her listen to its reasoning:

> "Is it an unclean thing that I have made you pregnant?"

> She did not understand the words at first. It was though it had begun speaking a language she did not know.
>
> "You … what?"
>
> "I have made you pregnant with Joseph's child. I shouldn't have done it so soon, but I wanted to use his seed, not a print. I could not make you closely enough related to a child mixed from a print. And there's a limit to how long I can keep sperm alive."
>
> She was staring at it, speechless. It was speaking as casually as though discussing the weather. She got up, would have backed away from it, but it caught her by both wrists.
>
> She made a violent effort to break away, realized at once that she could not break its grip. "You said—" She ran out of breath and had to start again. "You said you wouldn't do this. You said—"
>
> "I said not until you were ready."
>
> "I'm not ready! I'll never be ready!" (Butler 2000, 246)

Again Nikanj justifies its invasion of Lilith's body with its supersensory knowledge and utilizes force in order to accomplish its goal, both echoing and alluding to its treatment of Joseph:

> "You'll have a daughter," it said. "And you are ready to be her mother. You could never have said so. Just as Joseph could never have invited me into his bed-no matter how much he wanted me there. Nothing about you but your words reject this child." (Butler 2000, 247)

In impregnating Lilith, Nikanj utilizes both its privileged information about Lilith's body, and force, in order to control the outcome of the situation.

2.3 "You know you must accept me or Ooan": Dichotomous "Choices" in *Lilith's Brood*

Only occasionally do the Oankali appear to offer a choice. This choice is always a very narrow one; either it is between two options, neither of which are very favorable for the human, or it is a statement disguised as a question, intended to lead the human to the option the Oankali prefer. Lilith asks why she cannot have Jdahya as her teacher. Jdahya and his mate, Tediin, ask Lilith a series of questions that confirm for Lilith that there is no real choice:

> "… [I]f you and Nikanj weren't supposed to be teaching each other, you would be learning from Kahgyaht."
>
> Lilith shuddered. "Good god," she whispered. And seconds later, "Why couldn't it be you?"
>
> "Ooloi generally handle the teaching of new species."
>
> "Why? If I have to be taught, I'd rather you did it."
>
> His head tentacles smoothed.
>
> "You like him or Kahguyaht?" Tediin asked. Her unpracticed English, acquired just from hearing others speak was much better than Lilith's Oankali.
>
> "No offense," Lilith said, "but I prefer Jdahya."
>
> "Good," Tediin said, her own head smooth, though Lilith did not understand why. "You like him or Nikanj?" (Butler 2000, 71-72)

Lilith admits that she prefers Nikanj, adding, "'You people are manipulative as hell, aren't you?'" (Butler 2000, 72).

Lilith is again forced to choose between two unfavorable options when she learns that, no matter her wishes, the Oankali intend to make changes to her brain that will result in enhanced memory and nearly effortless language learning. Lilith is against the idea, saying, "'[N]o part of me is more definitive of who I am than my brain'" (Butler 2000, 76). Nikanj convinces Lilith to submit to the changes by telling her that surprising her would be "wrong" (Butler 2000, 78-79). Rather than preventing its ooloi parent, or "ooan," Kahguyaht, from altering Lilith's brain, Nikanj says that it will not surprise her, but "'you must trust me or let Ooan surprise you when it's tired of waiting'" (Butler 2000, 79). Lilith confronts Nikanj's hypocrisy when it says, "'We were bred to work with you … We should be able to find ways through most of our differences.'" "'Coercion,'" Lilith replies with rancor. "'That's the way you've found'" (Butler 2000, 81-82). Lilith knows that when force or coercion do not work, the Oankali resort to subterfuge.

2.4 "It should have told you": Sex and Deceit in *Lilith's Brood*

The Oankali control access to the information humans have in a number of ways. The first way in which they alter human knowledge is to block human survivors' access to memories of being captured by the Oankali: "'Humans who were allowed to remember their rescue became uncontrollable,'" sometimes killing themselves or others (87). The other ways in which the Oankali control human knowledge are much more insidious and much less altruistic. By denying Lilith information about how Oankali bonds function, the Oankali trick Lilith into accepting Nikanj as her mate. First they pair her with Nikanj as her Oankali teacher while it is still a child. Lilith thinks of Nikanj as a child, "no more responsible for the thing that was to happen to the remnants of humanity than she was" (Butler 2000, 72). Nikanj also tells Lilith that it is not, and cannot be, aroused by her (Butler 2000, 82). Butler associates the changes Nikanj makes to Lilith's brain with sexual coercion and deceit; Lilith learns, for instance, that Nikanj's performance of the brain "surgery" has left her bound to Nikanj:

> There was a faint odor to the hand—oddly flowery. Lilith did not like it and drew back from it after a moment of looking.
>
> Kahguyaht retracted the hand so quickly that it seemed to vanish. It lowered the sensory arm. "Humans and Oankali tend to bond to one ooloi," it told her. "The bond is chemical and not strong in you now because of Nikanj's immaturity. That's why my scent makes you uncomfortable."
>
> "Nikanj didn't mention anything like that," she said suspiciously.
>
> "It healed your injuries. It improved your memory. It couldn't do those things without leaving its mark. It should have told you." (Butler 2000, 110)

Soon after, Lilith becomes Nikanj's sexual partner (Butler 2000, 242). Lilith's ooloi children are even more duplicitous than is Nikanj, for they are able to change their appearance at will to appear more attractive to humans (Butler 2000, 604 & 630). Lilith is eventually complicit in the deceit, deliberately withholding from her ooloi child's potential human mates that if they choose to stay with Jodahs through its metamorphosis, they will never be able to leave (Butler 2000, 659-660).

Oankali subterfuge depends upon limiting the humans' access to the truth and dampening their perceptions. Nikanj admits that its pairing with Lilith was entirely premeditated, saying, "'You were being prepared for me, Lilith. Adults believed you would be best paired with me during my subadult stage. Jdahya believed he could bring you to me without drugs, and he was right'" (Butler 2000, 186).

2.5 "It would have outsold any illegal drug": Oankali Sex as Pharmacon

Humans who pose too much of a threat are drugged. As Nikanj notes, "'We dull your natural fear of strangers and of difference. We keep you from injuring or killing us or yourselves. We teach you more pleasant things to do'" (Butler 2000, 191-192). Nikanj also admits that the Oankali drugged "'newly awakened Humans much more than was good for them … because we saw … that we were damaging Lilith and the others" who had not been drugged, making them the target of their own people because the other humans perceive undrugged humans as having submitted willingly, even eagerly, to the Oankali (Butler 2000, 300).

Sexual pleasure is also a powerful drug in Butler's work. Lilith observes the trap of Oankali seduction: "Nikanj could give her an intimacy with Joseph that was beyond ordinary human experience. And what it gave, it also experienced. This was what had captured Paul Titus … This, not sorrow over his losses or fear of a primitive Earth" (Butler 2000, 161). Lilith's partner, Joseph, says of sex with the Oankali, "If a thing like that could be bottled, it would have outsold any illegal drug on the market" (Butler 2000, 169). Lilith, too, is addicted, as she more or less admits when Joseph asks why she has allowed the Oankali to have sex with her:

> "To have changes made. The strength, the fast healing—" He stopped in front of her, faced her. "Is that all?" he demanded.
>
> She stared at him, seeing the accusation in his eyes, refusing to defend herself. "I liked it," she said softly. "Didn't you?" (Butler 2000, 169)

Lilith makes similar statements to Tino, her mate after Joseph, in describing to him his conditioning by Nikanj when he was young and the reason he is so drawn to the Oankali:

"'Nikanj touched you when you were too young to have any defenses. And what it gave you, you won't ever quite forget—or quite remember, unless you feel it again. You want it again. Don't you.' It was not a question" (Butler 2000, 294). Lilith is convinced of the power of the Oankali drug – the physical sensations that come with stimulation via the ooloi's sex organ – to win over most anyone who has felt it. This is why Akin, Lilith's construct son, says to one of the human resisters, "You wanted to [stay with the Oankali] … You still do" (Butler 2000, 363).

2.6 The Irresistibility of Alien Control

As Lilith says to another of the human resisters, "'We're all a little bit co-opted, at least as far as our individual ooloi are concerned'" (Butler 2000, 240). It is not *just* the ooloi sex, though; it is the way in which the humans first experience ooloi sex. It is the lack of information about what an ooloi's touch will do to them. It is that, knowing most humans would never agree to sexual contact of their own accord, the first contact is almost always forced or done under the guise of some other action. It is the juncture of desire, force, deceit, limited agency, and sex that has made the humans so malleable and integrated them into the folds of the Oankali.

3. Cognition, Coercion, Force, and Consent in *Fledgling*

The Ina can, and do, dominate humans just as the Oankali do. They also use their extrasensory abilities to choose and to influence their symbionts, resort to physical force when necessary, prevent early symbionts from knowing the Ina's identity, offer humans limited options so that the Ina can determine the outcome, and utilize drugs. In addition, the Ina's use of the drugs means that they can compel humans under their influence to answer questions, remember information, and perform tasks. As with the Oankali, there is a strong sexual component to the Ina's relationship with their symbionts. Indeed, most of the Ina have sexual intercourse of some form or another with most or all of their symbionts. Like the Oankali, the Ina are able to determine whether or not a human is likely to be receptive to their sexual advances.

3.1 "I didn't imagine that loneliness had a scent": The Ina's Extrasensory Perceptions

Even without her memory intact, Shori relies on scent to tell her which humans to approach and try to convert to her symbionts. Her first convert, Wright, smells "really interesting" (Butler 2005, 15). Shori meets Wright by chance, but she chooses Theodora more carefully; Theodora's "aloneness was good, somehow … I got the impression that no one had touched her in a long time" (Butler 2005, 30). Shori tells Theodora, "'[Y]ou smell open, wanting alone…. longing, needing.'" Theodora asks, "'Do you mean that I smelled lonely? … I didn't imagine that loneliness had a scent…. I am lonely'" (Butler 2005, 98). In at least some fashion, Shori's choice of Theodora is completely calculated. Another Ina discusses the concept of a "'good symbiont,'" and Shori's choice of Theodora, with Shori:

> "… [S]he loves you absolutely. She's exactly the kind of person I would expect to be able to resist one of us—older, educated, well-off—but she couldn't wait to get to you."
>
> "She was lonely," I said. (Butler 2005, 207)

Because Theodora is lonely, Shori knows that Theodora will want to join Shori, Wright, and Shori's Ina family: "'She'll want to come. She doesn't have to, but she'll want to'" (Butler 2005, 93).

Shori also listens to voices and other cues to determine whether or not a human is scared or lying (Butler 2005, 251-252). Moreover, and perhaps more importantly, the Ina can influence humans and other Ina through their scent, both unconsciously and deliberately (Butler 2005, 216 & 222).

For some reason, though – perhaps to emphasize lack of consent – Butler never writes a character that is converted or seduced through scent alone. Force is always integral to Oankali and Ina dominance.

3.2 "I lay down beside the woman and covered her mouth with my hand… I held on to her with my other arm": Forced Sexual Encounters in *Fledgling*

Even after Shori bites Wright for the first time, she has to take his hand and forcibly keep it between hers while he tries to shake her off. Wright shakes Shori so vigorously that he lifts her "into the air a little." He continues to attempt to get away, but Shori is determined: "I didn't let go." Eventually Wright stops struggling (Butler 2005, 17). This pattern follows for all of Shori's symbionts and other humans that she bites (Butler 2005, 49 & 117). When Shori first bites Theodora, the force Shori uses is intense and extended:

> I lay down beside the woman and covered her mouth with my hand as she woke. I held on to her with my other arm and both my legs as she began to struggle. Once I was sure of my hold on hr, I bit into her neck. She struggled wildly at first, tried to bite me, tried to scream. But after I had fed for a few seconds, she stopped struggling. I held her a little longer, to be sure she was subdued; then, when she gave no more trouble, I let her go." (Butler 2005, 31)

Shori is stronger than all of her symbionts, including Wright (Butler 2005, 16). She uses force with abandon in each first bite.

3.3 "I can't leave you. I don't even want to leave you": Dichotomous "Choices" and Symbionts' inability to revoke consent in *Fledgling*

Once that contact is made, once a human is exposed to the Ina biochemical, it is difficult for that human to give up the pleasure. Wright tells Shori that it was impossible for him to choose to give her up, particularly given that she offered him the choice in a time of danger: "… [Y]ou think I could have just gone away and not come back? I had to leave you lying on the ground bleeding. You insisted on it. How could I not come back to make sure you were all right?'" (Butler 2005, 89). Wright points out the futility of Shori's offer when she asks him if he wants to leave:

> "Why bother to ask me that?" he demanded. "I can't leave you. I don't even want to leave you."
>
> "Then what do you want?"
>
> He sighed and shook his head. "I don't know. I know I wish I had driven past you on the road eleven nights ago and not stopped." (Butler 2005, 90).

Wright is so certain that he cannot have given consent once being exposed to Shori's drugged saliva that he wishes he had not met her at all. But save for the children of symbionts, no potential symbiont is ever offered the choice to be or not be a symbiont *before* exposure to the drug.

3.4 "I never really had a chance. I didn't have any idea what I was getting into": Ina Omission and Deceit

Wright says that humans let the Ina take them over "'because we have no choice. By the time we realize what's happened to us, it's too late.'" Brook, another symbiont, counters Wright: "'It's not usually that way … Iosif told me what would happen if I accepted him, that I would become addicted and need him. That I would have to obey. That if he died, I might die'" (Butler 2005, 167). Martin Harrison, however, disagrees with Brook's more generous assessment:

> "It doesn't seem to matter to most humans what our lives were before we met you. You bite us, and that's all it takes…. He bit me, and after that I never really had a chance. I didn't have any idea what I was getting into…. I wasn't physically addicted. No pain, no sickness. But psychologically … Well, I couldn't forget it. I wanted it like crazy." (Butler 2005, 210)

Though Brook suggests Shori's deceit is unintentional and "'probably because of her memory loss,'" and Wright says Shori has "'shown herself to be a weirdly ethical little thing most of the time'" (Butler 2005, 168), Shori is deceitful in other ways. In converting Theodora, perhaps anticipating Theodora's negative reaction to her skin color and apparent youth (Butler 2005, 95), Shori deliberately prevents Theodora from seeing her (Butler 2005, 31 & 94). Shori is deceitful repeatedly and intentionally in order to ascertain her symbionts' addiction and compliance.

3.5 "What I told them to do, they would try to do, once I had taken their blood": Pheromones, Biochemical Influences, and Sex in *Fledgling*

Deception is more closely tied to addiction in *Fledgling*. The limitations on human agency, also, are much more inextricably linked to biochemical drug addiction in *Fledgling* than in *Lilith's Brood*. Indeed, in *Fledgling*¸ the drug is more powerful. The drug's consequences for humans are more powerful as well. Not only can the addiction lead to death for human symbionts that lose their Ina, the biochemical affects *any* human who is bitten even once. For this reason an otherwise unaffected human can be led to give an Ina money or goods, divulge to an Ina privileged information, and even fight other symbionts in his or her family. Human symbionts must follow all orders given to them by their own Ina, and it is literally impossible for them to forget an order. Shori is aware of the power of her venom: "What I told them to do, they would try to do, once I had taken their blood" (Butler 2005, 110). The sexual pleasure inherent in the bite is also literally compelling. After first being bitten, Wright says he isn't sure he should allow Shori to do it again. Immediately after, he says, "'Shit, you can do it right now if you want to'" (Butler 2005, 24). The humans Shori bites ask, even beg, her to do it again (Butler 2005, 58 & 180), equating the experience of pleasure and addiction to cocaine (Butler 2005, 187). The humans truly feel that they need continued exposure to Ina venom, and the Ina take advantage.

3.6 "'[T]reat your people well'": Ina Control and (Lack of) Responsibility

Advanced sensory awareness, combined with superior strength and addictive venom, means that Shori, and the other Ina, can ostensibly have complete control over humans. Some of the Ina respect humans as autonomous beings, to an extent, as when Shori's father Iosif cautions her to be fair:

> "...[T]reat your people well, Shori. Let them see that you trust them and let them solve their own problems, make their own decisions. Do that and they will willingly commit their lives to you. Bully them, control them out of fear or malice or just for your own convenience, and after a while, you'll have to spend all your time thinking for them, controlling them, and stifling their resentment." (Butler 2005, 79)

Though he counsels her to be fair, Iosif sees the Ina as humans' "'more gifted cousin'" (Butler 2005, 73). Other Ina regard humans as no more than "tools;" weapons for murdering Ina or other symbionts (Butler 2005, 284-285). Even Shori admits, to one of her symbionts, that she "'won't always ask'" (Butler 2005, 289). For the Ina, asking for and receiving consent is an option, not a necessity.

4. Kant on the Ethics of Consent

In order to better articulate the implicit and explicit associations between the ethics of consent in Butler's work and rape culture, women's reproductive rights, and bioethics, a brief discussion of Kant's The Formula of the End in Itself follows. Kant is particularly relevant because Kant is clear on why deceit and coercion on the one hand, and consent on the other, are mutually exclusive. Kant is also clear that it is the lack of consent and of treating a person as an end in themselves that makes any particular course of action acceptable or not, rather than any products of that action, whether the products are for good or for ill.

4.1 Kant's The Formula of the End in Itself

Kant writes that one person should never treat themselves or another person as *merely* a means:

> [T]he human being, and in general every rational being, *exists* as end in itself, *not merely as means* to the discretionary use of this or that will, but in all its actions, those directed toward itself as well as those directed toward other rational beings, it must always *at the same time* be considered as an *end."* ([1785] 2002, 45)

Onora O'Neill extrapolates that "[t]o use someone as a *mere means* is to involve them in a scheme of action to which they could not in principle consent. Such situations include deceit:

> One person may make a promise to another with every intention of breaking it. If the promise is accepted, then the person to whom it was given must be ignorant of what the promisor's intention (maxim) really is.... Successful false promising depends on deceiving the person to whom the promise is made about what one's real maxim is. And since the person who is deceived doesn't know that real maxim, he or she can't in principle consent to his or her part in the proposed scheme of action. (1980, 287)

A second situation in which consent is impossible, O'Neill elaborates, is when coercion is involved. For instance, "[i]f a rich and powerful person threatens a debtor with bankruptcy unless he or she joins in some scheme, then the creditor's intention is to coerce; and the debtor, if coerced, cannot consent" (1980, 287).

4.2 Kant on Morality and "Rational Beings"

Kant also notes that not only humans are subject to these maxims against using a person as a mere means; rather, moral laws are applicable not only to human beings ([1785] 2002, 21), but all "rational" forms of life ([1785] 2002, 21 & 49). Thus, the Oankali and Ina are responsible for their treatment of human beings, and should be held accountable.

4.3 Humans as *Mere Means* in *Lilith's Brood* and *Fledgling*

Through unequal power dynamics, limiting the humans' options, deceit, force, and drugs/sex, the Oankali and Ina consistently treat the humans as a mere means rather than as ends in themselves. Because the humans are subjected to multiple and sustained constraints on their agency, they are unable to give consent. The other aspects of the relationship – better health, long lives, communal living, etc. – have very little or no bearing upon the morality of the Oankali's and Ina's actions because the humans did not enter into the beneficial aspects of the relationship with prior knowledge or willingness.

5. Rape Culture, Women's Reproductive Rights, and Bioethics

There are startling similarities between the Oankali's and Ina's treatment of the humans and the discourse of rape culture, women's reproductive rights, and bioethics. Like the humans in *Lilith's Brood* and *Fledgling,* victims of rape, women seeking abortions, and others in medical care situations, are often subject to reduced agency. They are told that their perceptions are inaccurate and/or that someone in a position of power has a greater access to the truth. They are forced into situations to which they do not want to and/or cannot give consent. They are presented with only a limited range of options, and they are tricked or drugged. What follows is a discussion of how *Lilith's Brood* and *Fledgling* apply to a discourse of rape and rape culture.

5.1 Rape and Rape Culture in *Lilith's Brood* and *Fledgling*

One must differentiate between the way Butler's work eroticizes and even romanticizes lack of consent in sexual intercourse, and actual rape culture. In other ways, however, there are a number of correlates between lack of consent in Butler's work and the ways in which lack of consent is discussed in other venues. Three of the Oankali and Ina strategies bear most closely upon the dynamics of a discussion of rape culture through *Lilith's Brood* and *Fledgling*: force, coercion, and drugs. The ways in which the Oankali and the Ina take advantage of humans are strikingly like the ways in which rape victims are first raped and then blamed as if though they entered into the sexual intercourse willingly. In both *Lilith's Brood* and in *Fledgling*, sexual relationships are initiated through force or through deceit. Joseph, for instance, is laid down on the bed against his will by Nikanj. Lilith's first sexual encounter with Nikanj is under the guise that it is *only* making changes to her brain.[2] Tino is too young to have defenses against the Oankali's sexual

[2] Nikanj notes before this encounter that it is too young to make the experience pleasurable for Lilith. Frances Bonner posits that the omission of the first physically pleasurable, and purely sexual, activity between Lilith and Nikanj is telling: "Butler presents this scene [the sexual encounter between Joseph and Nikanj] with the male rather than the female human and indeed does not show us the scene where Nikanj first rapes/seduces Lilith at all. It occurs between the first and second sections of *Dawn* and is not even

coercion. Shori surprises Wright when she first bites him, and literally has to hold down Theodora.

A 2000 study by the U.S. Department of Justice found that fifty-four percent of the rape victims surveyed were under 18 at the time of the assault (Thoennes and Tjaden 2000), and sexual assault is often a feature of domestic abuse ("Victims and Perpetrators" 2010). The humans in *Lilith's Brood* and *Fledgling,* like actual rape victims, are often in vulnerable situations.

The humans in *Lilith's Brood* and *Fledgling* are also drugged, as is often the case with actual rape victims, and particularly those who experienced sexual assault while attending university. The National Institute of Justice Campus Sexual Assault survey (2007) found that though "[m]ore women experienced forced sexual assault before college than during," it was more common for college students to be sexually assaulted while incapacitated, whether through drugs or alcohol (Krebs et al. 2008, 5-1 – 5-3). Men who participated in this study and who had been sexually assaulted reported higher rates of incapacitated sexual assault than forced sexual assault (Krebs et al. 2008, 5-5). There is a strong correlation between Butler's human characters and these victims of sexual assault because, as with human "trade partners" and "symbionts," the college students reported being "unable to provide consent" (Krebs et al. 2008, 5-2).

Another way in which Butler's human characters are like rape victims is their inability to say "no" and have that statement respected as truth. This inability to effectively dissent is particularly true of the humans in *Lilith's Brood.* Joseph says he doesn't want to have sex with the Oankali. He tells Nikanj, "'Let go of me.'" Nikanj says, "'Be grateful, Joe. I'm not going to let go of you.'" Nikanj explains, "'Your body said one thing. Your words said another'" (Butler 2000, 190).

The disconnect between a rape victim's words and their other actions is often a feature of the discourse surrounding rape and rape culture. Linda A. Bell notes that "judges and jurors might look at a perpetrator's intention, worrying about the injustice of punishing one who … really believed his victim was consenting (Bell 1993, 176). Posters from Project Unbreakable, in which rape victims write what their rapists said to them just before or after the assault, include these statements: "I know you want it." "You know you want it." "We both know you don't really mean it when you say no." "'You said no,

recalled in memory…. With Lilith there to assure the reader that the sexual experience is pleasurable and something she is all too willing to engage in herself, rape more easily masquerades as seduction. Her own first encounter, devoid of any such commentary, would be difficult to present convincingly as a desirable experience" (1990).

but your body told me yes'" (Koehler 2013). A consent infographic circulating in various forms and originally based on a tumblr post discusses consent and the lack thereof in detail:

> NO means NO.
>
> STOP means NO.
>
> TURNING AWAY means NO.
>
> PUSHING AWAY means NO.
>
> 'LEAVE ME ALONE' MEANS NO.
>
> PASSED OUT means NO.
>
> 'I'M NOT READY' means NO.
>
> 'I DON'T FEEL LIKE IT' means NO.
>
> INTOXICATED means NO. ("_______ means _______" 2014)

The statements and actions above, and the ones which follow on the original, all represent the statement "no." However, rape is legally defined in most states as sexual intercourse "'when the offender purposely compels the other person to submit by force or threat of force'" (Tuerkheimer). Both Joseph's and Lilith's first sexual encounters with the Oankali could be considered as rape by such standards, as could Wright's and Theodora's with Shori. All of the sexual encounters in *Lilith's Brood* and *Fledgling* are preceded by force. However, like a rape victim who has been drugged or is otherwise intoxicated beforehand or during, the humans in *Lilith's Brood* and *Fledgling* would not have the ability to give consent, even if offered the opportunity. Butler's work lends nuance to portrayals of sexual intercourse that do not truly involve consent.

The Oankali and Ina, like rapists, fail to see their victims as ends in themselves. The Oankali and Ina emphasize the symbiotic nature of the relationship they have with humans, as well as their own needs – for the Oankali, to "'trade … [o]ur genetic material for yours'" (Butler 2000, 40) and for the Ina, to "find several people to take blood from" (Butler 2005, 21). As Lilith says to Joseph, of Nikanj, "'I doubt whether it really cares what either of us wants'" (Butler 2000, 170). The Oankali and Ina, as Michele

M. Moody-Adams writes of rapists and rape apologists, do not "respect the integrity and *separateness* of the victim" (1990, 203). Similarly, Lilith's pregnancy occurs because of the lack of respect for Lilith as an end in herself, and the way her pregnancy is made known to her functions as an illumination of women's reproductive rights.

5.2 Women's Reproductive Rights in *Lilith's Brood*

Lilith's pregnancy is forced upon her. Years later, Nikanj still insists that the pregnancy is what Lilith wanted:

> Tino turned toward Lilith but spoke to Nikanj. "Did you make her pregnant against her will?"
>
> "Against one part of her will, yes," Nikanj admitted. "She had wanted a child with Joseph, but he was dead…. In the first children, I gave Lilith what she wanted but could not ask for." (Butler 2000, 300)

When Lilith thanks Nikanj for making Akin appear to be human, Nikanj says, "'You have never thanked me before…. And I think you go on loving them even when they change'" (Butler 2000, 254).

In both instances, Nikanj insinuates that Lilith needed only to get used to the idea of being pregnant (with an alien). Such an insinuation is not so different from the coercive tactics of those who are against abortion, or pro-lifers. In particular, Nikanj's action resembles the pervasive laws in the United States that require women to receive counseling, wait anywhere from 12-72 hours ("An Overview of Abortion Laws" 2015), and view - or at least be offered the chance to view - an ultrasound before undergoing an abortion procedure ("Requirements for Ultrasound" 2015). In Canada, there are no such laws regarding restrictions on abortion; however, there are approximately 200 (as opposed to 4000 in the United States) Crisis Pregnancy Centres which also aim to prevent abortions through the use of misinformation and coercion (Khandaker 2014). Moreover, two bills introduced in Canada in recent years – Bill C-484 and Bill C-510 – also relied on the premise that women would realize the value of pregnancy and motherhood either during or after the pregnancy, with Bill C-484 suggesting that "women are incapable of understanding the mother-child relationship they are forfeiting until they see their child born" (Davies 2009 13) and Bill C-510 "protecting against coerced abortion but not coerced childbirth" (Davies 2011 1).

Such tactics, like Nikanj's in impregnating Lilith without her knowledge and then using force and coercion in order to gain her cooperation, again do not respect the right of a woman to be an end in herself. Such tactics ignore that, like Lilith, many women who are pregnant have not freely given their consent to participate in a sexual relationship (Bell 1993, 21 & 26) or had the opportunity to prevent conception (Bell 1993, 26) in the first place. They are also indicative of the general tendency for institutions of medicine, whose representatives are largely male, to make decisions for women and to "coerce women into seeing an unwanted pregnancy through" (Sherwin 1989, 66-67). The deceitful and coercive tactics and acts used by anti-abortion activists also suggest that, just as the Oankali and Ina believe of humans, women who seek an abortion are incapable of reasoning and acting on their own.

5.3 Bioethics and Consent in *Lilith's Brood* and *Fledgling*

Butler makes the association between medical care, especially neurological changes, and sexual coercion and deceit, when Lilith learns that Nikanj's changes to her brain have also resulted in a sexual connection with and addiction to Nikanj. That association persists more subtly throughout *Lilith's Brood* and also *Fledgling,* and is underscored by the Oankali and Ina assumption that humans must be led. While the link between sexual coercion and medical care is plausible, it is also tenuous, though performing certain examinations without informed consent could be considered "extreme battery" – for instance, in the case of a patient who unwillingly undergoes a testicular cancer exam (Eyal 2011, 10).

The link between the overall coercion of humans in *Lilith's Brood* and *Fledgling* and the rising concern with bioethics, however, is more substantial. Nir Eyal observes that concern for informed consent as a predominant feature of bioethics grew in the twentieth century, especially "in medical research on human subjects … in reaction to abuses" (2011, 1). A patient who has given informed consent must be competent, as well as be aware of and understand the treatment procedures (Eyal 2011, 3). Eyal posits that informed consent is important in order to avoid abusive contact (2011, 11-12) and domination (2011, 15), as well as preserve trust (2011, 12-15), self-ownership (2011, 14-15), and personal integrity (2011, 15-17). For a patient to truly give informed consent, interactions with the physician must be free of "[l]ies about pertinent matters," "non-lying deceit," and "partial disclosure" (Eyal 2011, 19-20). Informed consent practices must also be free of "coercion" (Eyal 2011, 24-25); "undue inducement," or an offer "that is alluring to the point that it clouds rational judgment" (Eyal 2011, 25); and "so-called no choice situations" (Wertheimer 1987 qtd. in Eyal 2011, 26). Since "medicine is rife with potential to become hierarchical, given the utter dependency of patients and research participants on physicians" (Levine 1988 qtd. in Eyal 2011, 15), informed consent is necessary.

In *Lilith's Brood* and *Fledgling*, the Oankali and Ina lie, trick humans through various means, and disclose only partial information or no information at all. Coercion and undue influence both occur also, though it is mainly inducement via biochemical addiction that spurs the humans to continue to serve the Oankali and the Ina. Lilith's choice between brain alternations made by either Nikanj or Kahguyaht can be seen as a "no choice situation." In all cases, the humans in these novels are in situations where they have very little or no agency.

6. Conclusion

Lilith's Brood and *Fledgling* do not ever clearly equate the Oankali and Ina's treatment of humans with rape and rape culture, women's reproductive rights, and bioethics. Rather, *Lilith's Brood* and *Fledgling* trace associations between rape, women's reproductive rights, and bioethics to show the ways in which constraints on agency via access to privileged information, force, deceit, limited choice, and drugs can result in nearly complete control of a subject. What is clear is that the humans in the two novels do not have the right to choose, any more than do rape victims, women coerced into initiating or sustaining a pregnancy, and many medical patients. They are not respected as ends in themselves, and as such, cannot give consent. More thorough examinations of Butler's work promise to continue to illuminate the ethics of consent, contribute to a growing body of scholarship on agency in Butler's work, and initiate nuanced but responsible public discourse on rape, women's reproductive rights, and bioethics.

References

“______ means ______.” Last modified April 8, 2014, accessed August 29, 2015, http://persephoneholly.tumblr.com/post/82100763066/means#_=_.

Ackerman, Erin M. Pryor. “Becoming and Belonging.” The Productivity of Pleasures and Desires in Octavia Butler’s Xenogenesis Trilogy.” *Extrapolation* 49 (1): 24-43.

“An Overview of Abortion Laws.” *Guttmacher Institute.* Last modified August 1, 2015, accessed August 29, 2015, http://www.guttmacher.org/statecenter/spibs/spib_OAL.pdf.

Bast, Florian. 2010. “’I won’t always ask’: Complicating Agency in Octavia Butler’s *Fledgling.*” *Current Objectives of Postgraduate American Studies* 11. http://copas.uni-regensburg.de/article/view/128/152.

Bonner, Frances. 1990. “Difference and Desire, Slavery and Seduction: Octavia Butler’s *Xenogenesis.*” *Foundation* 48: 50-92.

Butler, Octavia E. 2005. *Fledgling.* New York: Seven Stories Press.

Butler, Octavia E. 2000. *Lilith’s Brood.* New York: Grand Central Publishing.

Butler, Octavia E. 1981. “The Science Fiction of Octavia Butler.” Carolyn S. Davidson. *Sagala* 2 (1): 35.

Bell, Linda A. 1993. *Rethinking Ethics in the Midst of Violence: A Feminist Approach to Freedom.* Lanham, MD: Rowman and Littlefield Publishers, Inc.

Davies, Cara. “Stereotyping and the new Women-protective Antiabortion Movement” (paper presented at the 2009 John and Mary Yaremko Forum on Multiculturalism and Human Rights: Student Symposium on Women’s Human Rights, Toronto, Ontario, March 6, 2009).

Davies, Cara. 2011. “Protecting Women or Peddling Stereotypes? Bill C-510 and the Influence of the Woman-protective Anti-abortion Movement.” *Journal of Law and Equality*: 1-26.

Eyal, Nir. 2011. “Informed Consent.” *Stanford Encyclopedia of Philosophy.* Last modified September 20, 2011, accessed August 2, 2015, http://plato.stanford.edu/entries/informed-consent/.

Fink, Marty. 2010. “AIDS Vampires: Reimagining Illness in Octavia Butler’s *Fledgling.*” *Science Fiction Studies* 37 (3): 416-432.

Govan, Sandra Y. 1986. “Homage to Tradition: Octavia Butler Renovates the Historical Novel.” *MELUS* 13 (1/2): 79-96.

Kant, Immanuel. (1785) 2002. *Groundwork for the Metaphysics of Morals.* Edited and translated by Allen W. Wood. New Haven: Yale University Press.

Khandaker, Tamara. 2013. "Phony Abortion Clinics in Canada Are Scaring Women with Lies." *Vice.com.* Last modified June 26, 2013, accessed August 28, 2015, http://www.vice.com/en_ca/read/i-went-to-a-phony-abortion-clinic-in-toronto.

Koehler, Sezin. "From the Mouths of Rapists: The Lyrics of Robin Thicke's 'Blurred Lines, Using Images from Project Unbreakable, an online photo essay exhibit." Last modified September 19, 2013, accessed August 30, 2015, http://www.psmag.com/books-and-culture/mouths-rapists-lyrics-robin-thickes-blurred-lines-66569.

Krebs, Christopher P. et al. 2008. "The Campus Sexual Assault (CSA) Study." *U.S. Department of Justice.* Last modified October 2007, accessed September 18, 2015. https://www.ncjrs.gov/pdffiles1/nij/grants/221153.pdf.

Moody-Adams, Michele M. 1991. "Gender and the Complexity of Moral Voices." In *Feminist Ethics,* edited by Claudia Card. Lawrence, KS: University Press of Kansas.

O'Neill, Onora. 1980. "The Moral Perplexities of Famine Relief." In *Matters of Life and Death,* edited by Tom Regan, 260-298. Philadelphia: Temple University Press.

Pickens, Theri. 2014. "'You're Supposed to be a Tall, Handsome, Fully Grown White Man': Theorizing Race, Gender, and Disability in Octavia Butler's *Fledgling.*" *Journal of Literary and Cultural Disability Studies* 8 (1): 33-48.

"Requirements for Ultrasound." *Guttmacher Institute.* Last modified August 1, 2015, accessed August 29, 2015, http://www.guttmacher.org/statecenter/spibs/spib_RFU.pdf.

Sherwin, Susan. 1989. "Feminist and medical ethics: two different approaches to contextual ethics." *Hypatia* 4 (2): 57-72.

Thoennes Nancy and Patricia Tjaden. 2000. "Full Report of the Prevalence, Incidence, and Consequences of Violence Against Women: Findings From the National Violence Against Women Survey." Washington, DC: U.S. Department of Justice, National Institute of Justice. Last modified July 2000, accessed August 29, 2015, https://www.ncjrs.gov/pdffiles1/nij/181867.pdf.

Tuerkheimer, Deborah. "We preach 'no means no' for sex, but that's not what the law says." *The Guardian.* Last modified January 12, 2014, accessed August 29, 2015, http://www.theguardian.com/commentisfree/2014/jan/12/rape-definition-use-of-force.

"Victims and Perpetrators." *National Institute of Justice*. Last modified October 26, 2010, accessed August 30, 2015, http://www.nij.gov/topics/crime/rape-sexual-violence/pages/victims-perpetrators.aspx.

Journal of
Cognition
and
Neuroethics

Evolution and Neuroethics in the *Hyperion Cantos*

Brendan Shea
Rochester Community and Technical College

Biography

Brendan Shea, PhD, is Instructor of Philosophy at Rochester Community and Technical College and a Resident Fellow at the Minnesota Center for Philosophy of Science. His current research focuses on issues related to the philosophy of science and applied ethics, and to philosophical methodology more generally. He also has an abiding interest in the interconnections between science fiction and philosophy.

Acknowledgements

I'd like to thank the participants of the 2015 "The Work of Cognition and Neuroethics in Science Fiction" conference for their many helpful comments.

Publication Details

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

Citation

Shea, Brendan. 2015. "Evolution and Neuroethics in the *Hyperion Cantos*." *Journal of Cognition and Neuroethics* 3 (3): 139–162.

Evolution and Neuroethics in the *Hyperion Cantos*

Brendan Shea

Abstract

In this article, I use science-fiction scenarios drawn from Dan Simmons' "Hyperion Cantos" (*Hyperion*, *The Fall of Hyperion*, *Endymion*, *The Rise of Endymion*) to explore a cluster of issues related to the evolutionary history and neural bases of human moral cognition, and the moral desirability of improving our ability to make moral decisions by techniques of neuroengineering. I begin by sketching a picture of what recent research can teach us about the character of human moral psychology, with a particular emphasis on highlighting the importance of our evolutionary background as social mammals. I then consider how the moral psychology of intelligent machines might differ from our own, and argue that the differences would depend on the extent to which their evolutionary background resembled our own. I offer two very different case studies—the "Technocore AIs" that have evolved from early, parasitic computer programs, and the mysterious "Shrike," who travels backward through time. I close by looking at the character of Aenea, a messianic figure that is a joint descendant of humans and machines. I argue that while the sort of "moral enhancement" she represents is far beyond the scope of either contemporary neuroscience or artificial intelligence research, it nevertheless represents a worthwhile goal.

Keywords

Evolutionary Ethics, Machine Ethics, Moral Cognition, Moral Enhancement, Neuroethics, Science Fiction

While serious work on moral psychology goes all the way back to Aristotle and Hume, and preliminary investigations of the evolutionary bases of morality can be found in Darwin's *Descent of Man,* it is only in the last few decades that these two projects have begun to converge in meaningful, productive ways. Modern classics such as E.O. Wilson's *Sociobiology: The New Synthesis* (1975) and Richard Dawkins' *The Selfish Gene* (1976) have led to an ever-increasing amount of research on the evolutionary pressures that shaped human moral behavior. During this same period, neuroscience has made impressive gains in its ability to locate (and in some cases, to manipulate) moral responses in the brains of both humans and non-human animals. Recent years have seen a number of prominent attempts to tie these strands together to provide both descriptive accounts of why and how human morality has developed as it has, and normative proposals based on these accounts.[1]

[1] Some prominent examples include Wright (1994), Dennett (1996; 2006), Pinker (1997), Sober and Wilson

In this paper, I'll be investigating some of the key themes of this recent research in the context of Dan Simmons's "Hyperion Cantos," a series of four books that appeared between 1989 and 1997. I have two major goals. First, I'll be exploring the extent to which human moral norms are the product of our unique evolutionary heritage and to what extent we could reasonably expect intelligent beings with *different* evolutionary pasts to share them. Second, I'll consider how (and whether) the results of this descriptive moral project bear on the normative project of improving human moral behavior. With this in mind, I'll conclude by considering the potential for so-called "moral enhancement" by technological means. I will argue that such actions would, subject to certain caveats, be both permissible and desirable.

1. Background to the Hyperion Cantos

Simmons' Hyperion Cantos consists of two pairs of books: *Hyperion* (1989) and *The Fall of Hyperion* (1990)*,* and *Endymion* (1996) and *The Rise of Endymion* (1997)*,* all of which are set in the distant future. When the series begins, humanity has already colonized a large number of worlds, and developed a technologically advanced society with the help of a highly evolved group of artificial intelligences called the Technocore (or "Core"). The Canto's plot is driven by the conflicts between human civilization and the Technocore, and between both groups and a breakaway group of humans known as the "Ousters," who are distinguished by their extensive use of bioengineering techniques to adapt their bodies to harsh, non-earthlike environments.

The first two books in the Cantos take their names from John Keats's unfinished poems "Hyperion" (Keats 1977, 283–307) and "The Fall of Hyperion: A Dream" (1977, 435–449) which deal with the conflict between the Greek Titans (including the sun god Hyperion) and their Olympian Children, who will eventually replace them.[2] The titles of

(1999), Singer (2000; 2011b), Greene (2001; 2013), Preston and de Wall (2002), Haidt (2001; 2012), de Waal (2009; Waal 2014), Churchland (2011), Harris (2011), and Wilson (2013).

[2] Keats' "Hyperion" is presented as a third-person narrative focusing on the successive replacement of old gods by new ones (Chronos/Saturn replaces Uranus/Caelus, and is himself replaced by Jove/Uranus). When the poem begins, Hyperion is the only Titan who remains in power. While the "Fall of Hyperion" incorporates substantial text from the original poem, the context is much different: in this case, it one aspect of a first-person "dream," which deals much more explicitly with the subjects such as the value of art, its relationship to death, and so on. One of the main characters of Simmon's Hyperion Canon—the woman "Moneta" who travels backward through time with the Shrike—shares her name with the goddess of memory who plays a major role in Keats' "The Fall of Hyperion." Keats abandoned both poems before finishing them.

the final two books refer to Keats' long poem "Endymion," which tells the story of a love affair between a human shepherd and the goddess "Cynthia," or Artemis (Keats 1977, 106–217). The books are filled with numerous references to both Keats and his work, and the plot is set in motion by the actions of a half-human, half-AI John Keats "cybrid" that has been designed by elements of the Technocore to have the memories and values of the historical Keats.

Like Keats' original poems, the books explore questions such as: "What, if anything, will come after humanity as it exists now?" and "What role, if any, do things like love, empathy, and art play in improving human life?" In the first two books, a group of seven pilgrims goes on a quest to save humanity from a rumored Ouster invasion. This quest takes them to a planet called Hyperion, and places them in conflict with a horrifying being called "The Shrike," which moves backward in time, is seemingly invulnerable, and whose main goal seems to be to capture various beings to torture on its "tree of thorns." The imminent invasion is eventually revealed to be a ploy by (certain elements) of the Technocore, who want to destroy humanity in order to prevent the evolution of a highly empathetic human "God" in the far future, which will compete with the (much less empathetic) AI "Ultimate Intelligence." The next set of two books (set several hundred years further into the future) deal with Aenea, the daughter of the John Keats' clone and one of the pilgrims. Aenea is a messianic figure who represents the next "stage" of both human and AI moral evolution, and she eventually resolves the conflicts that arise from the Core's and humanity's divergent moral norms. The Shrike again plays a major role in Aenea's quest, though in this case it is generally helpful, presumably because the events of the first Hyperion books have altered the circumstances leading to its eventual creation.

2. Parental Care as the Basis for Mammalian Morals

Before turning to the vexing questions of how non-human moral systems might work, or what this means for the possibility of improving human moral cognition, it will be helpful to briefly review some key findings of recent neuroscience and moral psychology as they relate to human moral cognition. In many cases, these findings are both surprising and counterintuitive, and they will play a key role in later parts of the argument.

According to one dominant tradition descended from thinkers such as Plato, Kant, and Freud, humans' capacity for moral and altruistic behavior is tied tightly to humans' capacity to use dispassionate and impartial *reason* to overrule their baser drives and instincts. This view is exemplified, for example, in the influential social contract theory of Thomas Hobbes (1994), who sees morality as a sort of *agreement* among rational agents to "play by the rules" for mutual advantage.

In recent years, however, research in areas such as social neuroscience, cognitive psychology, and zoology has cast doubt on this "reason-centric" picture of human moral behavior. A variety of studies (J. D. Greene et al. 2001; J. Greene and Haidt 2002; Haidt 2001; Haidt 2007) strongly suggest that many "prototypical" human moral judgments are driven mainly by automatic, intuitive emotional processes and not by higher-order cognitive processes.[3] This picture coheres well with recent research on primatology (Flack and de Waal 2000; Warneken et al. 2007; Waal 2009; Waal 2009), which has suggested that close analogues of human "morality" can be found in non-human primates such as chimps and bonobos, who presumably lack the capacity for explicit, reason-based moral theorizing. Finally, recent research (Insel 2010; Churchland 2011) on the neurology of ethical decision-making has begun to identify the specific brain areas and neuropeptides (such as oxytocin and argine vasotocin) involved in ethical decision-making, and provided some promising suggestions on how our ability to care about others, and to take action on their behalf, might have evolved.

In more practical terms, this research suggests that humans' moral-decision making is at least as strongly shaped by our long evolutionary past as social mammals as by our ancestors' (far more recent and limited) experience with explicit moral theorizing and argumentation. Here, some examples from the Hyperion Cantos will help clarify things. To begin with, let's consider maternal and paternal care, which plausibly form the evolutionary "bedrock" of mammals' more generalized ability to form caring relationships. In the Hyperion Cantos, this sort of ground-level concern for offspring is exemplified by the pilgrim Sol Weintrub, a Jewish ethicist whose daughter Rachel has been infected by a "Merlin's sickness" that has caused her to age backward through time, and to slowly lose all memories of everything that has happened to her. Sol, unsurprisingly, identifies so strongly with Rachel's loss that it seems almost physically painful to him, and he is willing to do anything (giving up his job, spending all of his savings, voyaging across the universe) in an attempt to save her.

[3] The role played by moral theorizing, or by higher-order reasoning more generally, has been a matter of some debate. Haidt (2001; 2012) argues that the content of moral decisions is determined almost entirely by immediate, automatic processes. By contrast, some prominent utilitarians (J. D. Greene et al. 2001; Singer 2005; J. Greene 2013) argue that this is true only of deontological (or non-utilitarian) moral decisions, and have pointed to fMRI data showing that utilitarian judgements are associated with relatively less emotional engagement.

Sol is specifically appalled by recurrent dreams in which the Shrike appears and demands that he hand over his daughter Rachel as a "sacrifice" to save humanity from destruction. This sort of Abrahamic sacrifice, it seems to Sol, is one that is deeply immoral, and one that cannot be squared with a truly "human" morality. While he eventually consents to it (when an adult Rachel appears to him in a dream and requests this), this does not resolve the underlying ethical tension. When considered from the lights of an impartial morality, Sol's actions verge on the incomprehensible—after all, the best evidence he has suggests both that (1) it is *very* unlikely that Rachel can be saved and (2) that the results of *not* sacrificing Rachel to the Shrike may be catastrophic. Given this, it seems that a purely "rational" father (even one who cares deeply about his daughter) would choose to sacrifice Rachel's small chance of salvation in order to save humanity (including both himself and his daughter) from almost certain destruction a short time later. However, Sol's actions fit well with the emerging picture of mammalian moral decision-making sketched above, according to which threats to one's children as processed (quite literally) in the same way as threats to one's own life.[4]

3. Expanding the Circle of Concern

The human ability to care about others is not constrained to parents and children, of course. Like most fictional works, the Hyperion Cantos contains numerous examples of self-sacrifice and heroism performed on the behalf of romantic partners, friends, and even strangers. To begin with, let's consider romantic love. In the first two books, Brawne Lamia repeatedly risks her life to save the cybrid Keats, with whom she eventually becomes romantically involved, and even agrees to carry his memories in a "neural shunt" after his physical body is destroyed by the Technocore. In the last two books, Raul Endymion serves first as a young Aenea's protector, and then later as the mature Aenea's

[4] Sol's dilemma here bears some resemblance to Foot's (1967) and Thomson's (1976; 1985; 2008) famous "trolley" cases, in which a person is offered a choice between two courses of action, one of which will lead to a single person's death, and one of which will lead to a larger number of deaths. In recent years, these scenarios have played a key role in investigations into the psychology and neuroscience of moral decision-making (J. D. Greene et al. 2001; Cushman, Young, and Hauser 2006; Koenigs et al. 2007; Uhlmann et al. 2009; Liao et al. 2012). People's judgements (including those of moral "experts") in these sorts of cases have been found to be highly context-sensitive, and to vary according to cognitive load, the order in which the cases are presented, the amount of direct physical force applied in the killing, the race of the victim, and many other factors. The apparent inconsistency, combined with peoples' difficulty in justifying their judgments (specifically in those cases where they let the greater number die, and violate utilitarian norms), strongly suggest that "automatic" processes play a significant role.

spouse. In all of these relationships, just as was the case in the parental relationship between Sol and Rachel, threats to one's mate are experienced neurologically in much the same way as threats to *oneself*. This fits well with recent research on pair-bonding in both rodents and primates (Insel and Hulihan 1995; Young and Zuoxin Wang 2004; Liu and Wang 2003; Smith et al. 2010), which suggests that many of the same neural mechanisms at work in paternal care also play important roles in enabling some mammals to form long-term relationships, and in grounding their capacity to *care* deeply about what happens to their mate.

Going beyond parental and romantic relationships, the ability to form well-functioning social groups among *non-relatives* has been crucial to the success of most primates, including both modern humans and our ancestors. So, for example, the seven Hyperion pilgrims of the first two books come from radically different cultural, religious, and even biological backgrounds. Through the process of sharing their unique stories, however, they begin to "cohere" into a tight-knit group in which individuals are willing to make considerable sacrifices for their companions, and even for "humanity" in general. This ability of radically different humans to "come together" in the face of adversity is widespread, and it is something like a "staple" of standard science fiction stories (and of fiction more generally). Again, while these relationships are not *identical* to parental and romantic relationships, they rest on quite similar cognitive and affective capacities, such as the ability to experience another's pain and suffering as "one's own," and to be motivated to *do* something about it. It should not be surprising then, to discover that evolution has recruited many of the same neural mechanisms involved in grounding parental and pair-bonding relationships to allow our brains to understand, and care about, those who are *not* related to us (Immordino-Yang et al. 2009; Zak, Stanton, and Ahmadi 2007; Iacoboni 2009; Shamay-Tsoory 2011).

This research suggests that the human brain's capacity to care about the well-being of others has its evolutionary origins in first, the sorts of neural mechanisms relied upon by vertebrates to maintain their *own* bodily integrity, and more recently, in the specific extension of these mechanisms in mammals to allow for extended maternal care of offspring. These same mechanisms have then been recruited to allow for things such as paternal care, concern about mates, and so on. Finally, in many social mammals (including humans), these mechanisms have been further modified to allow concern for those who are neither mates for kin, but are member of one's "group." This final step of extending caring to non-relatives and non-mates, of course, plausibly calls for a somewhat different evolutionary explanation. In particular, where the extension of caring behavior toward offspring may largely be a matter of kin selection, explaining the broader concern of social mammals toward other group members might involve also involve appeals to reciprocal altruism, group selection, or both.[5]

While it is undeniable that groups whose members care about one another provide concrete advantages to individuals in terms of things such as personal safety and resource allocation,[6] there is also the risk that selfish individuals may take advantage of the concern of others, and act to benefit themselves at others' expense. It should be no surprise, that both humans and their primate relatives regularly punish cheaters and rule-breakers, even when doing so represents a significant personal cost. In the Hyperion Cantos, this characteristic of human moral psychology is best exemplified by the character of the Consul, the one-time Hegemony-appointed ruler of Hyperion who (before the books begin) has betrayed the Hegemony by agreeing to serve as an "agent" of the Ousters. Importantly, the Consul is motivated not by self-interest, but by a desires for *punishment, revenge* and *justice.* In the Consul's story, he reveals that his grandparents had been rebels against the Hegemony, who had conquered (and then ruthlessly exploited) their home world. Later, when he discovers the Core's malignant intentions for humanity, he attempts to strike back at it by betraying the Ousters as well, and prematurely triggering a device that releases the Shrike (whose actions the Core can neither predict nor control) from the "Time Tombs."

While the Counsel comes to regret aspects of both his actions and the motives that drove them, they rely upon important, and widely shared, aspects of human moral psychology. In particular, the Consul, like many other humans, shows that he is willing to punish "cheaters" and "rule breakers" (such as the Hegemony and the Core) even when doing so is *not* in his own self-interest, no matter how widely this is construed. According to a number of recent studies (Fehr and Gächter 2002; Boyd et al. 2003; Barclay 2006; Marlowe et al. 2008), it is precisely the presence of "altruistic punishers" (and the deterrence they provide for potential rule breakers) such as the Consul that allowed early

[5] The respective role of kin selection, reciprocal altruism, and group selection in explaining human sociality, of course, a matter of some debate. Dawkins (1976) and Wilson (1975) famously reject group selection, and provide accounts of human sociality and altruism grounded in kin selection and reciprocal altruism. Sober and Wilson (1999) and Wilson (2013) by contrast, argue that group selection also played a significant role. While this debate is clearly of independent interest, my thesis here does not depend on any particular resolution.

[6] Some recent research suggests that the human brain's larger capacity for social cognition may have given human groups significant advantages over those of Neanderthals, specifically in areas such as the ability to trade for exotic goods, and to maintain innovations across generations (Pearce, Stringer, and Dunbar 2013).

humans to form social groups significantly larger than those of their primate ancestors and relatives.

4. Some Complications: "In Groups" and "Out Groups"

So far, I have focused on the on the ways in which human morality can be seen as a natural outgrowth from our origins as social mammals. In particular, I've looked to the Hyperion Cantos to illustrate more general points about our abilities to understand and care about offspring, romantic partners, and selected others within our communities in much the same way that we care about our *own* well-being. These capacities served our ancestors well, as they helped to ensure stable, tight-knit communities where members "looked out" for one another by doing things such as providing resources to those who need them (such as the young or sick), defending the defenseless, and enforcing prohibitions against those community members who "cheat."

There is, however, a dark side to human morality as well, both in its tendencies to disproportionately punish norm violations by group members, and by its seeming disregard for those who are *not* members. These tendencies are prominently on display throughout the Hyperion Cantos, just as they are in the real world. The secular, pseudo-democratic Hegemony of the first two books, for instance, has regularly committed genocide against non-human species that it worries may someday evolve to challenge humanity. The Catholic "Pax" government which takes the Hegemony's place in the second two books is equally vicious, and murders or kidnaps whole populations of non-Christians in an attempt to keep Aenea's "virus" from spreading and destroying the immortality-granting "Cruciform" technology on which Pax power is based. Both the Hegemony and the Pax regularly engage in bloody, offensive wars against the "unnatural" Ousters, who they think have forfeited their humanity by virtue of their use of their "unnatural" bioengineering techniques on their own bodies to adapt to life in harsh environments.

While it is tempting to think that these undesirable aspects of human psychology are fundamentally opposed to our evolved capacity for moral reasoning, and of having their origin in entirely different motivations and mechanisms, there are good reasons to think this is mistaken. Instead, recent work has suggested that many of the same neural processes that ground our strong, intuitive concern for "in-group" members, and to justly and proportionately punish wrongdoers, may also predispose us (at least is some cases) to violence against out-group members, and to disproportionately and unjustly punish violations of "purity" (Tybur et al. 2013; Haidt 2012; Dreu et al. 2011; Hammond and Axelrod 2006; Dreu et al. 2010; Haidt and Graham 2007; Hodson and Costello 2007). Some authors have suggested that it was precisely the demands of intergroup conflict and war that provided the evolutionary impetus for primates' (and humans') evolved ability to form coalitions, and their attendant in-group morality, in the first place (Hammond and Axelrod 2006; Tooby and Cosmides 2010). Others (Fiske, Rai, and Pinker 2014) argue that morally-motivated violence remains a wide-spread, and often underappreciated, social problem. This all suggests that, insofar as we want to count things like empathy, compassion, and a concern for justice, as core elements of "human nature," we must *also* count such things as racism, religious discrimination, interpersonal violence, and our general tendency to think of outgroup members as being less worthy of concern than are the members our own group.

On reflection, the hypothesis that there is a close relationship between dedication to an "in-group" and hatred of an "outgroup" should not strike us as implausible. Consider, for example, institutions such as the military or organized religion, both of which play major roles in the Hyperion Cantos. On the one hand, these highly disciplined, hierarchical, and uniquely human institutions can help extend the boundaries of the "in-group" membership far beyond what is possible for any non-human primates. Colonel Kassad, for instance, manages to overcome his background as an orphaned, impoverished member of a religious minority to rise to a high position within the Hegemony military, while Father de Soya overcomes a similarly impoverished background to become a leader in the Pax's "new" Catholicism. On the other hand, as both characters painfully discover, the coherence of these institutions depends crucially on the institutions ability to enforce strict obedience to (seemingly arbitrary) norms, and on the existence of an "outgroup" against which to define themselves. While the cultivation of in-group loyalty is not in itself bad, it does mean that they, like all human institutions, are vulnerable to moral perversion. When this happens—the military goes to war against the Ousters, the Pax attacks religious minorities—it can be very difficult for those within these institutions to both recognize these undesirable changes and to arrest them.

While there is not room here to explore the relationship between evolution, morality, and religion in anything like the detail it deserves, Simmons' picture of a post-cataclysmic revival of "traditional" religious beliefs and organizations in the Endymion books fits with some current thinking about the relationship between religion and ethics. More specifically, while it seems highly implausible that religion plays much of a role in determining the *content* of human moral norms (since these norms clearly predate religious belief, and can survive its absence), it may help "unify" large, disparate groups by allowing the members of these groups to "extend" their moral trust and concern outside the boundaries of their small community. Moreover, unlike "rival" solutions to the problem of group harmony (such as those provided by well-functioning liberal democracies), religion is relatively "simple," and does not require many institutional prerequisites to establish or maintain (Dennett 2006; Churchland 2011; Fukuyama 2012; Waal 2014; Norenzayan 2014).

5. Machine Ethics: Some Possible Scenarios

So far, we have focused primarily on human morality. I have suggested that many features of human morality, such as our willingness to make sacrifices for our children, mates, friends, and other "in group" members are tightly tied to our evolutionary history as social mammals. The survival of our mammalian and primate ancestors depended crucially on their abilities to protect and educate their children, and to cooperate effectively with non-relatives to do things such as hunt or engage in inter-group aggression. In order to accomplish this, evolution recruited brain areas originally designed to detect threats to *self* to register and respond to threats to selected *others*. It also enabled them to detect cheaters and rule-breakers, and motivated them to punish, even at a personal cost. Our moral capacities thus rest on both our cognitive ability to understand and predict the behavior of others, and the affective inclination to respond appropriately.

If this picture is correct, then we have some reason to think that intelligent biological life-forms on other planets might well have evolved moral norms similar to humans, at least if their ancestors had to spend significant amounts of time nurturing their young, and had to live within social groups. These beings would, like us, care about other members of their group, but be prone to distrust and dislike beings "outside" this group. While such beings are relatively rare with the Hyperion Cantos, the few examples given (such as the evolved dolphins of Maui Covenant) seem to fit this description.

In the context of Hyperion Cantos, the far more interesting question concerns the potential character of machine ethics. Citing Thomas Ray's early work on the "Tierra" model of artificial life (1991; 1993), Aenea suggests that the advanced AIs of the Technocore had their evolutionary origins as *parasites.* In particular, the ancestral, human-made programs of the Technocore AIs were forced to compete for limited CPU power in order to replicate themselves. The winning strategy in these early days, at least according to Aenea, was to function as "parasites" that shed the (costly) ability to "self-replicate," and instead hijacked *other* programs' code to replicate themselves. This led to a spiraling sort of "hyper-parasitism," where the evolving AIs became better and better at using the resources of both other AIs and their human hosts in order to replicate themselves. Where social mammals had invested their resources in a joint project of caring and defending their vulnerable offspring, which were their genes' only "hope for the future," each individual AIs within the Technocore had the potential for immortality, so long as it could continually *self-evolve* (largely by incorporating bits of destroyed competitors, or by capturing new computing resources from their human "hosts"). By the time the Hyperion Cantos begin, the Technocore AIs have perfected this strategy, and have begun directly using human neurons for their own processing purposes.

Unsurprisingly, the ethics of a highly evolved parasite look very different from those of social mammals. In particular, where the humans of the Hyperion Canon find it relatively easy to form and maintain tight-knit groups, the self-interested Core AIs are forced to navigate a world of rationally negotiated, short-lived alliances, and in which the primary strategy for gaining resources is to exploit their human "partners." While some of the Core AIs (the "Ultimates") have devoted themselves to the creation of an Ultimate Intelligence that will someday subsume everything within itself, a larger number (the Stables and the Volatiles) seek to maintain their existence as individuals, either by continuing to serve as parasites on humans, or by destroying them and finding alternate mediums. Insofar as this picture seems plausible, we should be wary of *assuming* that properties such as intelligence and moral concern for others will necessarily co-evolve, at least in the context of machines.[7]

As some of the Core AIs eventually come to recognize, however, this way of life is hugely inefficient, since it requires individuals to devote *massive* amounts of resources merely to maintain the status quo. It is partially for this reason that they create the Keats cybrid, which is a "machine mind" that realizes valuable parts of human morality, including the capacity for empathy, while still retaining a Core AI's ability to impartially focus on the "big picture" as opposed to one's narrow "in group." While the actions of this cybrid (and its child, Aenea) eventually lead to the destruction of the Technocore,

[7] Axelrod (1981; 1984), among many others, has argued that generally altruistic strategies (such as "tit-for-tat") carry significant advantages over purely "selfish" ones, at least in certain sorts of competitive games (such as Prisoner's Dilemma). This provides at least some reason to think that, were the Core AIs entirely cut off from the resources to be gained from their human "hosts," their descendants might *eventually* gravitate toward "altruistic" or "nice" ways of dealing with one another, at least in many contexts. However, there is little reason to think that machine moral psychology would mirror the norms of human moral psychology, given their very different evolutionary heritages. In any case, this future eventuality would plausibly be of little consolation to the humans immediately endangered by the Core's actions.

the book strongly suggests that those silicon-based intelligences that *do* survive will now evolve on the model of Keats, and have effectively "overcome" their parasitic past.

While the parasitic ethics of the Core AIs are distinctively non-human and non-mammalian, they are nevertheless capable of certain types of altruistic and cooperative behavior. The Ultimates, for instance, are perfectly willing to sacrifice their individual "lives" to help "give birth" to the Ultimate Intelligence, while the Volatiles and Stables are capable of forming symbiotic relationships with both each other and humans. Such behaviors can be easily explained, for example, by the sorts of reciprocal-altruism-based accounts of group cooperation often used by evolutionary biologists to explain group dynamics for a wide variety of organisms. Core AIs, insofar as they want the help of other beings to further their own goals, have at least *some* reason to keep their promises and to avoid obvious "cheating." However, they appear to lack the other sorts of mechanisms (such as altruistic punishment or concern for kin), which form the bedrock for humans' abilities to genuinely "care" about the well-being of others.

The time-reversed Shrike, by contrast, is an intelligent being that lacks even this primitive moral base. While it is clearly a *future* product of joint human and Core evolution, its changing motives throughout the Hyperion Cantos strongly suggest that the precise circumstances of its evolutionary past are underdetermined by present events. The Shrike appears to be, in the words of Daniel Dennett, an evolutionary "good trick," which represents a good "solution" to a problem that will arise in a wide variety of (future) environments. That is, it seems that *some* group in the future will create the Shrike in an effort to fulfill *some* purpose; however, which group (and which purpose) will do this isn't determined. In the first two books, the Shrike appears to have been created by, and to be serving the will of, the future Core UI in its war against the empathetic human "God" that may be a product of future evolution.[8] In the final two books, by contrast, it appears to be serving Aenea's purposes, though it is clearly beyond her (or anyone else's) control.

The Shrike, unlike the Core AIs, might be a physically (and perhaps even logically) *impossible* being. So why care about it? One reason is that the Shrike represents a sort of thought-experiment: What would it take to create an intelligent being that lacked

[8] One of the main characters of the Hyperion Cantos, Father Paul Duré, begins as an adherent of Pierre Teilhard de Chardin (1965), who had argued that God was an (inevitable) product of future evolution, and the books spend considerable time exploring variants of this view. However, the scenario described in the Hyperion Cantos does not fit with Teilhard's (highly contentious and unorthodox) claims regarding biological evolution, and the character Aenea at one point rejects these views as incomplete or inaccurate.

any recognizable moral code? The answer, the Hyperion Cantos suggest, is to create a being that lacks any determinate evolutionary past, that cannot engage in repeated social interactions of any type, and which is incapable of being harmed or destroyed. Under these conditions, and under no others, can such a being be imagined. Another reason for caring about the Shrike is that it, or something like it, may represent something like a dark counterpart to the sort of "desirable" moral evolution that Aenea represents. Like Aenea, it is a "hybrid" of human and machine; unlike Aenea, however, it is a being utterly stripped of even the most basic moral norms. The Shrike is thus a sort of warning to those who would place blind faith in future evolution to make our descendants "better" than we currently are.

6. Engineering Ethically Better Beings

The events of the Hyperion Cantos suggest that the key to "overcoming" the shortcomings in the dominant human and AI ways of moral-decision making is to somehow expand the scope of the "in group" to include absolutely *all* sentient beings, regardless of how different their interests might be. The Keats cybrid, for example, represents a "machine" that can empathize with human suffering, while his daughter Aenea has a unique (and seemingly biologically-based) ability to cognitively and affectively empathize with *all* sentient beings who have ever lived. This idea—that moral progress requires "expanding the circle" of our moral concern, and of replacing our selective moral concern with a truly "impartial" empathy—is roughly consonant with evolution- and neuroscientific-based arguments for utilitarianism by Harris (2011), Singer (2011b), Greene (2013), and others.

But how can this be accomplished? One limited mechanism for doing this may involve artistic creativity. So, for example, the poet Martin Silenus (the purported "author" of the Cantos) appears to have indirect access to the thoughts and motivations of nearly all the major actors within the story, including the other Shrike pilgrims, the Ousters, the Hegemony and Pax leaders, and many others. This, of course, an exaggeration of the *actual* capacities of any real-life artist. Nevertheless, the ability of narratives to help "tie" disparate individuals together should not be underestimated, and recent research has suggested that reading narrative fiction can indeed enhance empathy (Mar and Oatley 2008; Kidd and Castano 2013).

While things such as narrative fiction, art, religion, and philosophy are clearly important first "steps" in broadening our moral horizons, the Hyperion Cantos suggests that these alone will not be enough, unless these things motivate us to take practical steps to *engineer* morally better beings. The Keats' cybrid, for example, is an engineering marvel that represents a radically different sort of moral being than the dominant Core mode of existence. If "artificial" life-forms on this model are to flourish in a world of limited resources, however, this means that the more "traditional" Core AIs that would compete for these resources will inevitably lose out (and perhaps even face extinction, as is suggested at the end of the Cantos). As Simmons recognizes, this is a conclusion which many of the Core AIs find highly unpalatable, and which they are willing to fight to stop.

This argument has conclusions that go beyond artificial intelligence, however. After all, if we find it morally acceptable to engineer morally better AIs by "pruning away" the morally outdated ones, we may need to consider doing the same things for *humans,* who (just like the Core AIs) are all too prone to making moral mistakes. And this is precisely what the Cantos suggests is necessary. Aenea is herself, after all, a sort of "engineering project" designed by elements of the Technocore and (perhaps) by other, highly evolved beings known only as the "Lions, Tigers, and Bears." More importantly, her "solution" to the problems presented by existing human institutions is in large part an *engineering* one. In virtue of her unique biology, she able to infect (willing) people with an "Aenea virus," that will (1) destroy the "cruciforms" the have rendered humans effectively immortal (and thus prevented death from doing its necessary work in evolutionary progress) and (2) allow humans a *vastly* increased ability to empathetically identify with other sentient beings. People who have been affected by Aenea's virus can, among other things, *literally* feel the pain of others they hurt, and are cognitively emotionally affected by the experiences of beings everywhere. While Aenea repeatedly argues that these biological changes are not *sufficient* for moral progress, she suggests that they may at least be *necessary.* It may simply be impossible, she suggests, for "traditional" humans to ever overcome their tendencies toward violence and selfishness.[9]

If Aenea is right, then we are morally *obligated* to engage in (voluntary) bio- and neuro-engineering projects aimed at "moral enhancement." A similar proposal (albeit in a

[9] The Aenea virus seems to grant those it infects immediate, phenomenological access to the pains, pleasures, and preferences of everyone else. This plausibly provides a strong psychological impetus for adopting a form of maximizing utilitarianism, according to which one's only (moral) duty is to maximize happiness (or preference satisfaction), regardless of whose happiness or satisfaction this is. One potential worry, raised both the character of Raul Endymion, and by prominent critics of utilitarianism (Williams 1973; Wolf 1982; Nagel 1989; Friedman 1991), is that this sort of "universal" and "impartial" concern is incompatible with having "integrity," or with engaging in the sorts of projects and relationships that make human life worthwhile. Aenea, in keeping with utilitarian responses to these objections (Railton 1984; Sosa 1993; Jackson 1991; Driver 2005; Singer 2011a) disagrees with this characterization of characterization.

very different context), has been defended by Persson and Savulescu (2008; 2012), who have argued that continuing *technological* process (in particular, in the realm of non-moral cognitive enhancements) represents a profound threat to the future of humanity, since it provides us with increasingly efficient and effective methods of self-destruction. While engineering changes on the scale of the Aenea virus are far beyond the scope of current methods, Douglas (2008) argues that we may soon be able to undertake more limited interventions, such as those aimed at reducing violent aggression or aversion toward other races.

There are, of course, a number of (potentially serious) worries about moral enhancement that would need to be considered it could be deployed, even supposing we had the technological means to do so. Harris (2011), for example, argues that pursuing moral enhancement is undesirable, at least if "moral enhancement" is understood to be distinct from cognitive enhancement more generally. While dealing with Harris's arguments in detail is beyond the scope of this article, I do not think that any of them amount to *in principle* arguments against moral enhancement, at least of the sort represented by the Aenea virus. So, for example, Harris objects to Douglas's proposal that racism (and other forms of harmful discrimination) could be combatted with neural enhancements aimed at diminishing the (often negative) *affective* reactions that humans experience when interacting with out-group members. Harris suggests that (1) there are other, less intrusive ways of diminishing the impact of racism (such as education) and (2) direct interference with the mechanisms that generate distrust and dislike of outsiders may "weaken kinship ties or other ties unconnected with race," as well as moral reactions more generally (2011, 105). This follows from the fact (noted earlier) that many of the same neural mechanisms involved in our (often negative) response to out-group members are crucial in enabling in-group cohesion.

Whatever the cogency of Harris's arguments when applied to Douglas's proposal, they do not apply the "Aenea" model of moral enhancement, which is primarily a *cognitive* enhancement, as opposed to an *affective* one. In particular, the Aenea virus functions not by directly intervening on a peoples' *reactions* to old experiences, but providing them with *new experiences* that allow them to "see" more directly the concerns of other people, in much the same way that they can see their *own* concerns. This, unlike the proposals that worry Harris, would not require direct interference with the brain's capacity to care about others, or to form attachments.

Another of Harris's arguments, however, may be more directly relevant to the Hyperion Cantos. Harris argues, contra Persson and Savulescu, that we should not delay or suspend research into (non-moral) cognitive enhancement technology, even in cases where these cognitive enhancements plausibly increase the power of individuals to do massive harm, and even when we do not yet have the capacity to engineer moral enhancements to help counteract these increased risks. An example here may help. Perrrson and Savulescu are worried that rapid increases in human cognitive capacity (specifically those brought about by neuroengineering) may lead to a situations where a single individual (perhaps because of malevolence or simple ignorance) can cause a significant amount of harm (for example, by using their enhanced abilities to design and utilize a new type of weapon). They argue that, insofar as it generally easier for an individual to cause massive harm than to cause a benefit of similar magnitude, we have some reason to refrain from pursuing such technologies, at least until research on moral enhancement catches up. Harris, in contrast to Perrrson and Savulescu, contends that there is no cogent argument for supposing *a priori* that future cognitive enhancements will *disproportionately* raise the risk of harm, when weighed against their potential benefits. Instead, the history of science provides some evidence to the contrary: while a wide variety of scientific research can and has been harnessed to inflict great harm (nuclear or biological weapons), this same research has also led to significant benefits for humanity (space travel, nuclear power, or antibiotics).

While the considerations raised by Harris are both significant and relevant, the scenario provided by the Hyperion Cantos provide some evidence for thinking that these sorts of arguments are not unlimited in scope. Consider, for example, the original technology that eventually gives "birth" to the Core AIs—a group of (relatively simple) computer programs that are exposed to evolutionary pressures that push them toward greater and greater cognitive capacities, capacities that can (when they reach the so-called "singularity") be used to consciously "self-engineer" further increases in these same capacities. In a scenario widely echoed in contemporary science fiction, the Core AIs eventually turn on their (less cognitively adept) human creators. One can, with a little effort, imagine similar doomsday scenarios resulting from the use of neuroengineering used to improve human intelligence.

The point here is not that the mere conceptual possibility of apocalypse-by-machine should lead us to suspend research into either artificial intelligence or human cognitive enhancement. As Harris cogently argues, to do so might mean forfeiting significant potential benefits. However, it does suggest—contra Harris—that it would a mistake to take "increased cognitive capacity" as being the *sole* target of our engineering efforts in these areas, at least if our aim is to increase human welfare. Instead, we should recognize (as the characters of Hyperion—both machine and human—eventually come to) the distinctive role that moral norms (and the related notions of *empathy* and *concern*) play in enabling a worthwhile existence, and consciously consider questions concerning such norms in our scientific efforts.

In the case of artificial intelligence, this may mean applying our knowledge of human moral cognition (both its evolutionary history and underlying neural mechanisms) in efforts to produce genuinely "social" and "moral" machines. This does not mean, however, that we can or should design machines to precisely mirror human moral norms. After all, as I've tried to suggest, these norms are far from perfect, and may *themselves* someday be targets for potential intervention. And indeed, it is not implausible to expect that these two research projects—the design of "moral machines" and potential techniques for human moral enhancement—are tied tightly to one another, and that discoveries in one area will contribute to a more comprehensive understanding the other.

7. Conclusion

The careful consideration of thought experiments has a long history within philosophical ethics, and the extension of this methodology to the scenarios provided by longer works of science fiction is a natural one. It holds particular promise for investigating questions regarding the potential evolutionary and neural underpinning of human moral cognition, and for examining in particular the extent to which our norms are the result of contingencies of our evolutionary heritage as social mammals. As I've tried to suggest here, answering these questions is of considerable practical, as well as theoretical, import, especially as we begin to seriously evaluate the prospects for designing "moral" machines and for developing techniques for human moral enhancement.

References

Axelrod, Robert. 1981. “The Emergence of Cooperation among Egoists.” *American Political Science Review* 75 (02): 306–18.

———. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Barclay, Pat. 2006. “Reputational Benefits for Altruistic Punishment.” *Evolution and Human Behavior* 27 (5): 325–44.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. “The Evolution of Altruistic Punishment.” *Proceedings of the National Academy of Sciences* 100 (6): 3531–35.

Chardin, Teilhard de, and Sir Julian Huxley. 1965. *The Phenomenon of Man*. Translated by Bernard Wall. 2nd edition. New York: Harper & Row/Harper Torch Book.

Churchland, Patricia S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton: Princeton University Press.

Cushman, Fiery, Liane Young, and Marc Hauser. 2006. “The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm.” *Psychological Science* 17 (12): 1082–89.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dennett, Daniel C. 1996. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster.

———. 2006. *Breaking the Spell: Religion as a Natural Phenomenon*. Reprint edition. New York: Penguin Books.

Douglas, Thomas. 2008. “Moral Enhancement.” *Journal of Applied Philosophy* 25 (3): 228–45. doi:10.1111/j.1468-5930.2008.00412.x.

Dreu, Carsten K. W. De, Lindred L. Greer, Michel J. J. Handgraaf, Shaul Shalvi, Gerben A. Van Kleef, Matthijs Baas, Femke S. Ten Velden, Eric Van Dijk, and Sander W. W. Feith. 2010. “The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans.” *Science* 328 (5984): 1408–11. doi:10.1126/science.1189047.

Dreu, Carsten K. W. De, Lindred L. Greer, Gerben A. Van Kleef, Shaul Shalvi, and Michel J. J. Handgraaf. 2011. “Oxytocin Promotes Human Ethnocentrism.” *Proceedings of the National Academy of Sciences* 108 (4): 1262–66. doi:10.1073/pnas.1015316108.

Driver, Julia. 2005. “Consequentialism and Feminist Ethics.” *Hypatia* 20 (4): 183–99.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415 (6868): 137–40.

Fiske, Alan Page, Tage Shakti Rai, and Steven Pinker. 2014. *Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships*. Cambridge: Cambridge University Press.

Flack, Jessica C., and Frans de Waal. 2000. "'Any Animal Whatever'. Darwinian Building Blocks of Morality in Monkeys and Apes." *Journal of Consciousness Studies* 7 (1-2): 1–29.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5.

Friedman, Marilyn. 1991. "The Practice of Partiality." *Ethics* 101 (4): 818–35.

Fukuyama, Francis. 2012. *The Origins of Political Order: From Prehuman Times to the French Revolution*. Reprint edition. New York, N.Y.: Farrar, Straus and Giroux.

Greene, Joshua. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: The Penguin Press.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–8. doi:10.1126/science.1062872.

Greene, Joshua, and Jonathan Haidt. 2002. "How (and Where) Does Moral Judgment Work?" *Trends in Cognitive Sciences* 6 (12): 517–23. doi:10.1016/S1364-6613(02)02011-9.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34. doi:10.1037/0033-295X.108.4.814.

———. 2007. "The New Synthesis in Moral Psychology." *Science* 316 (5827): 998–1002. doi:10.1126/science.1137651.

———. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Vintage.

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize." *Social Justice Research* 20 (1): 98–116.

Hammond, Ross A., and Robert Axelrod. 2006. "The Evolution of Ethnocentrism." *Journal of Conflict Resolution* 50 (6): 926–36.

Harris, John. 2011. “Moral Enhancement and Freedom.” *Bioethics* 25 (2): 102–11. doi:10.1111/j.1467-8519.2010.01854.x.

Harris, Sam. 2011. *The Moral Landscape: How Science Can Determine Human Values*. Reprint edition. New York: Free Press.

Hobbes, Thomas. 1994. *Leviathan: With Selected Variants from the Latin Edition of 1668*. Edited by Edwin Curley. Underlined, Notations edition. Indianapolis: Hackett Publishing Company.

Hodson, Gordon, and Kimberly Costello. 2007. “Interpersonal Disgust, Ideological Orientations, and Dehumanization as Predictors of Intergroup Attitudes.” *Psychological Science* 18 (8): 691–98.

Iacoboni, Marco. 2009. “Imitation, Empathy, and Mirror Neurons.” *Annual Review of Psychology* 60: 653–70.

Immordino-Yang, Mary Helen, Andrea McColl, Hanna Damasio, and Antonio Damasio. 2009. “Neural Correlates of Admiration and Compassion.” *Proceedings of the National Academy of Sciences* 106 (19): 8021–26. doi:10.1073/pnas.0810363106.

Insel, Thomas R. 2010. “The Challenge of Translation in Social Neuroscience: A Review of Oxytocin, Vasopressin, and Affiliative Behavior.” *Neuron* 65 (6): 768–79. doi:10.1016/j.neuron.2010.03.005.

Insel, Thomas R., and Terrence J. Hulihan. 1995. “A Gender-Specific Mechanism for Pair Bonding: Oxytocin and Partner Preference Formation in Monogamous Voles.” *Behavioral Neuroscience* 109 (4): 782.

Jackson, Frank. 1991. “Decision-Theoretic Consequentialism and the Nearest and Dearest Objection.” *Ethics* 101 (3): 461–82.

Keats, John. 1977. *John Keats: The Complete Poems*. Edited by John Barnard. 3rd edition. Harmondsworth, New York: Penguin Classics.

Kidd, David Comer, and Emanuele Castano. 2013. “Reading Literary Fiction Improves Theory of Mind.” *Science* 342 (6156): 377–80.

Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio. 2007. “Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements.” *Nature* 446 (7138): 908–11.

Liao, S. Matthew, Alex Wiegmann, Joshua Alexander, and Gerard Vong. 2012. “Putting the Trolley in Order: Experimental Philosophy and the Loop Case.” *Philosophical Psychology* 25 (5): 661–71.

Liu, Y., and Z. X. Wang. 2003. "Nucleus Accumbens Oxytocin and Dopamine Interact to Regulate Pair Bond Formation in Female Prairie Voles." *Neuroscience* 121 (3): 537–44.

Marlowe, Frank W., J. Colette Berbesque, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Jean Ensminger, et al. 2008. "More 'altruistic' Punishment in Larger Societies." *Proceedings of the Royal Society of London B: Biological Sciences* 275 (1634): 587–92.

Mar, Raymond A., and Keith Oatley. 2008. "The Function of Fiction Is the Abstraction and Simulation of Social Experience." *Perspectives on Psychological Science* 3 (3): 173–92.

Nagel, Thomas. 1989. *The View From Nowhere*. Reprint edition. New York, NY: Oxford University Press.

Norenzayan, Ara. 2014. "Does Religion Make People Moral?" *Behaviour* 151 (2/3): 365–84. doi:10.1163/1568539X-00003139.

Pearce, Eiluned, Chris Stringer, and R. I. M. Dunbar. 2013. "New Insights into Differences in Brain Organization between Neanderthals and Anatomically Modern Humans." *Proceedings of the Royal Society of London B: Biological Sciences* 280 (1758): 20130168. doi:10.1098/rspb.2013.0168.

Persson, Ingmar, and Julian Savulescu. 2008. "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity." *Journal of Applied Philosophy* 25 (3): 162–77. doi:10.1111/j.1468-5930.2008.00410.x.

———. 2012. *Unfit for the Future: The Need for Moral Enhancement*. New York: Oxford University Press.

Pinker, Steven. 1997. *How the Mind Works*. New York: W. W. Norton & Company.

Preston, Stephanie D., and Frans De Waal. 2002. "Empathy: Its Ultimate and Proximate Bases." *Behavioral and Brain Sciences* 25 (01): 1–20.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13 (2): 134–71.

Ray, Thomas S. 1991. "An Approach to the Synthesis of Life." In *Artificial Life II*, edited by C Langton, C Taylor, JD Farmer, and S Rasmussen, 371–408. Redwood City, CA: Addison-Wesley.

———. 1993. "An Evolutionary Approach to Synthetic Biology: Zen and the Art of Creating Life." *Artificial Life* 1 (1_2): 179–209.

Shamay-Tsoory, Simone G. 2011. “The Neural Bases for Empathy.” *The Neuroscientist* 17 (1): 18–24.

Simmons, Dan. 1989. *Hyperion*. New York: Doubleday.

———. 1990. *The Fall of Hyperion*. New York: Doubleday.

———. 1996. *Endymion*. London: Headline Book Publishing.

———. 1997. *The Rise of Endymion*. New York: Bantam Books.

Singer, Peter. 2000. *A Darwinian Left: Politics, Evolution, and Cooperation*. New Haven: Yale University Press.

———. 2005. “Ethics and Intuitions.” *Journal of Ethics* 9 (3/4): 331–52. doi:10.1007/s10892-005-3508-y.

———. 2011a. *Practical Ethics*. 3 edition. New York: Cambridge University Press.

———. 2011b. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.

Smith, Adam S., Anders Agmo, Andrew K. Birnie, and Jeffrey A. French. 2010. “Manipulation of the Oxytocin System Alters Social Behavior and Attraction in Pair-Bonding Primates, Callithrix Penicillata.” *Hormones and Behavior* 57 (2): 255–62.

Sober, Elliott, and David Sloan Wilson. 1999. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Sosa, David. 1993. “Consequences of Consequentialism.” *Mind*, New Series, 102 (405): 101–22.

Thomson, Judith Jarvis. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–17.

———. 1985. “Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases.” *Yale Law Journal* 94 (6): 1395–1415.

———. 2008. “Turning the Trolley.” *Philosophy & Public Affairs* 36 (4): 359–74.

Tooby, John, and Leda Cosmides. 2010. “Groups in Mind: The Coalitional Roots of War and Morality.” *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, 91–234.

Tybur, Joshua M., Debra Lieberman, Robert Kurzban, and Peter DeScioli. 2013. “Disgust: Evolved Function and Structure.” *Psychological Review* 120 (1): 65.

Uhlmann, Eric L., David A. Pizarro, David Tannenbaum, and Peter H. Ditto. 2009. “The Motivated Use of Moral Principles.” *Judgment and Decision Making* 4 (6).

Waal, Frans de. 2009. *Primates and Philosophers: How Morality Evolved: How Morality Evolved*. Princeton: Princeton University Press.

———. 2014. *The Bonobo and the Atheist: In Search of Humanism Among the Primates*. New York: W. W. Norton & Company.

Warneken, Felix, Brian Hare, Alicia P. Melis, Daniel Hanus, and Michael Tomasello. 2007. “Spontaneous Altruism by Chimpanzees and Young Children.” *PLoS Biology* 5 (7): e184. doi:10.1371/journal.pbio.0050184.

Williams, Bernard. 1973. “A Critique of Utilitarianism.” In *Utilitarianism: For and Against*, by Bernard Williams and J. J. C. Smart. Cambridge, UK: Cambridge University Press.

Wilson, Edward O. 1975. *Sociobiology: The New Synthesis*. Cambridge, Mass: Harvard University Press.

———. 2013. *The Social Conquest of Earth*. New York: Liveright.

Wolf, Susan. 1982. “Moral Saints.” *The Journal of Philosophy* 79 (8): 419–39. doi:10.2307/2026228.

Wright, Robert. 1994. *The Moral Animal: Evolutionary Psychology and Everyday Life*. New York: Vintage Books.

Young, Larry J., and Zuoxin Wang. 2004. “The Neurobiology of Pair Bonding.” *Nature Neuroscience* 7 (10): 1048–54. doi:10.1038/nn1327.

Zak, Paul J., Angela A. Stanton, and Sheila Ahmadi. 2007. “Oxytocin Increases Generosity in Humans: e1128.” *PLoS One* 2 (11). doi:10.1371/journal.pone.0001128.

cognethic.org