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.