Review Paper

Autonomy Rebuilt: Rethinking Traditional Ethics towards a Comprehensive Account of Autonomous Moral Agency

Jeffrey Benjamin White

Korea Advanced Institute of Science & Technology, Korea
**corresponding author*: jeffreywhitephd@gmx.com

Abstract

Autonomous agency is complex, bound up as it is with moral agency. And, moral agency is anything but clear. Confronted with many unanswered questions, researchers often operate under two distinct notions of autonomy, one associated with human and another with artificial agents. This lack of uniformity is theoretically unappealing, impedes progress on both forms of agency, and its constructive resolution is the focus of this paper. First, we review Kant's account of autonomous agency, and then turn to some contemporary analyses in which this robust understanding of autonomy is reduced to suit artificial applications. From this reduction, we review some contemporary approaches to understanding autonomy, thereby opening a way back to a comprehensive account of agency. And finally, we integrate the results of this discussion into a model of autonomous agency that can serve both as a platform for testing theories of moral decision and action, and as a framework for engineering and evaluating autonomous agents and agency.

Keywords: Autonomous agent, artificial intelligence, moral decision and action

1. Introduction

Autonomous agency is complex, bound up as it is with moral agency. Indeed, "a moral agent is necessarily an autonomous agent."(Smithers, 1997, page 95) And, moral agency is anything but clear, bound up as it is with things like freewill, responsibility, intention, conscience, personal identity and selfhood. Confronted with so many un-answered questions, researchers often operate under two distinct notions of autonomy, one associated with human and another with artificial agents. This lack of uniformity is theoretically unappealing, impedes progress on both forms of agency, and its constructive resolution is the focus of this paper.

Towards this end, Anthony Beavers (2012) suggests that the "hard problem" in morality lies in "rearranging" the landscape of traditional moral concepts so that solutions to problems in engineering artificial moral agents (AMAs) present themselves. The alternative on his account is the possible end of ethics, "ethical nihilism," with traditional moral concepts such as conscience and autonomy replaced only by the hollow objective determination of an agent's position in a chain of efficient causation. Meanwhile, Wendell Wallach (2010) suspects that the lack of progress in ethics is due to a preoccupation with isolable moral faculties, rather than "recognizing that moral acumen emerges from a host of cognitive mechanisms" and that "all of those considerations either merge into a composite feeling or conflict in ways that prompt the need for further attention and reflection," with the moral agent necessarily functioning as an "integrated being."(page 249) Wallach thereby calls for a "comprehensive" account of moral agency, one that can serve as "a platform for testing the accuracy or viability of theories regarding the manner in which humans arrive at satisfactory decisions and act in ways that minimize harms."(page 248)

It is my suspicion that Wallach's demands can be met through something like Beavers' means. The conceptual resources necessary for a comprehensive account of autonomous moral agency are available in traditional ethics, but have been hidden behind conventional interpretations and summarily established conceptions of human relative to artificial agency. The present paper attempts some moral landscaping to stop the erosion of ethics into transactional recordkeeping. First, it reviews some analyses in which autonomy is reduced, inviting ethical nihilism. Then, it clears the way to a comprehensive account of agency. Finally, it constructs such an account from traditional materials, resulting in a model of autonomous agency that is not only integrated, but integrative, and that can serve both as a platform for testing theories of moral decision and action, and as a framework for engineering and evaluating autonomous agents and agency.

2. Recognizing distinctions

Etymologically, the term "autonomous" is ancient Greek, with ''auto'' meaning self, and ''nomos'' meaning law. Originally, it applied to societies, cities, and states, which were considered autonomous when their members lived according to custom and convention specific to their common environment, thereby creating their own laws,

rather than having laws externally imposed. Autonomy thus means "self-governing."

Immanuel Kant developed this original notion of autonomy in terms of individual moral agency, with the model for the autonomous agent being "the political sovereign not subject to any outside authority, who has the power to enact law," and autonomous thereby meaning "self-sovereign."(Reath, 2006, page 122) From here, Kant specified that each is not only able to create and to act from laws of his own creation, to be "rational," but "to pass judgment upon himself and his own actions" from the ideal vantage point of a "kingdom of ends," an ideal arrived at through the exercise of moral duty, for every "man" "to make mankind in general his end,"(Kant, 1780, page 26) meaning that every rational agent should identify its own interests with the "kingdom of ends" in the mode of the moral equivalent of the political sovereign. Famously, this Kantian agent is guided by a single principle, the categorical imperative, one form of which commands an agent "Never to employ himself or others as a mean, but always as an end in himself,"(Kant, 1796, page 37) with "end in himself" meaning self-sovereign, and so qualitatively equivalent with the agent, itself.(see Kant, 1788, page 89)

Autonomy thus requires that the autonomous moral agent be free from selfish material desires, thereby embodying virtue worthy of "reverence," i.e. deserving of the respect and admiration cum emulation deserving of a beneficent king. "Autonomy is therefore the ground of the dignity of humanity, and also of every other intelligent nature whatsoever."(Kant, 1796, page 39)

However, it is exactly this degree of autonomy that is not afforded artificial agents in their very conception. Consider Ronald Arkin's 2009 text *Governing lethal behavior in autonomous robots*, with the obvious concern, how one "governs" "autonomous" agents, an equally obvious oxymoron. On this account, robot "autonomy" is limited self-direction toward goals of external origin within a human command hierarchy, i.e. serving as means to another's ends. After all, self-legislating warrior-robots acting to preserve dignity dignity, rather than blindly following orders to maim and murder, run counter to the intractable role that Arkin presumes violence playing in the press of history. We will have more to say about this presumption, and what it means to our conception of autonomy, in a moment. Regardless, on such account, AMAs are better understood as AAAs, artificial *amoral* agents, with robot autonomy rather rendered as *not*-autonomy, at all.

A similar reduction is effected by Michael Arbib (2005). On his assay, humans enjoy the dignity of self-determination with each "finding his or her own path in which work, play, personal relations, family, and so on can be chosen and balanced in a way that grows out of the subject's experience rather than being imposed by others," while for a robot "the sense is of a machine that has considerable control over its sensory inputs and the ability to choose actions based on an adaptive set of criteria rather than too rigidly predesigned a program."(Arbib, 2005, page 371) With artificial agents cast as objects of purely external determination rather than moral subjects, this is also a characterization of *not*-autonomous amoral agency. Finally, the rigid distinction between human and artificial moral agency is articulated by Tom Ziemke (2008) in terms of a Kantian inspired distinction between the "phenomenal" and "noumenal," with the first ascribed and the latter emerging via autopoietic self-organization, and with robots ultimately lacking the material constitution necessary to emerge as autonomous in the full sense, *not*-autonomous and so amoral by default.

In each preceding case, researchers propose two classes of autonomy so different that it is difficult to trace them to same concept at all. This divide can be smoothed over by rendering differences in autonomy by degree, however. Prima facie, there are three degrees of autonomous agency applicable to artificial agents. One, as a direct extension of human agency, only. This is a machine on auto-pilot, for example a landmine or a BMW on cruise control. Two, as an indirect extension of human agency. This is the conception most common to artificial agents, that they will act according to interred rules fed top-down, whether categorical principles or conditional guidelines. Arkin's military robots serve as good examples here; as human soldiers follow codes of warfare, so should their machines. The third degree specifies autonomy in the fully sovereign sense, representing both the promise of continued research in artificial intelligence – a fully autonomous AMA – and the promise of traditional moral education – autonomous human agents (AHAs) thriving in a just world of their own creation.

James Moor's is perhaps the most influential graduated analysis of autonomous moral agency.(Moor, 2006, 2007) At the lowest level, any agent or artifact the actions of which have ethical consequences qualifies as an "ethical impact agent." Moor offers the replacement of human jockeys with robotic jockeys in Qatar as an example here, whereby humans were freed from torturous servitude by machines unable to suffer similarly. One level higher, "implicit ethical agents" are morally significant by design. Moor's examples of such are spam-bots and airplane instruments that warn pilots of unsafe conditions, clearly degrees of currently realized ethical "agency," and still direct extensions of human agency. Moor's third type of ethical agent, the "explicit ethical agent," is able to identify morally salient information within specific contexts and to act according to appropriate principles. An indirect extension of human agency, Moor feels that this is the "paradigm case" of robot ethics, "philosophically interesting" and "practically important" while not too sophisticated to be realized. Finally, Moor's fourth type of ethical agent is the "fully ethical agent," by Moor's estimation not a level of agency likely to be realized in robots, representing self-sovereign agency characterized by three distinctly human characteristics - free will, consciousness, and intentionality - the engineering of which present serious problems.

Schermerhorn and Scheutz (2004) have also proposed a graduated classificatory schema. Theirs includes perceived

autonomy in the spirit of Ziemke's "phenomenal" autonomy. Their first degree of autonomy involves executing some function without direct human assistance.

Robotic jockeys qualify here, as would robotic vacuum sweepers. The second involves following human directives, without the need for step-by-step direction. Military missions would qualify as such directives, with this degree expressed by mission-capable agents. Schermerhorn and Scheutz's third level involves goal self-ascription and independent decision-making facilitated by self-reflective capacities over intentional states, corresponding to the fully ethical agent on Moor's hierarchy.

Finally, Schermerhorn and Scheutz point to a neglected aspect of autonomous agency, the perception and ascription of autonomy based on demonstrations of agency. For instance, in some situations, an autonomous agent will simply sweep the floor when put to that task, while in others it will stop sweeping to save the neighbor's cat from a burning barn, with this latter demonstration inviting an ascription of autonomy and the former, not.

The trouble here is that autonomy involves the context sensitive capacity to do the right things at the right times, and fully autonomous agents do not always appear that way, confounding any easy ascription on phenomenal bases.

Autonomy is "adjustable," and the demonstrated capacity to adjust the degree of autonomy that an agent expresses is essential to both autonomous agency and its ascription. Being a fully autonomous agent often involves ceding autonomy through "transfer-of-control," reflecting the fact that even fully autonomous agents pursue objectives of external derivation, e.g. as part of a team.(Pynadath et.al., 2002) Extending the context of team to include family, company, society, it becomes clear that most human action is externally determined, with original and on-going control over one's own "path in life" ceded well prior to birth and as a matter of course. In light of this fact, any distinction between human and artificial agent based on apparent source of guidance and goal may be misplaced. And this poses a real problem, not only for our ascription of autonomy to artificial agents, but for any conception of autonomous agency, at all.

For instance, consider that in the great team that is the military, both robots and humans are embedded within the same command hierarchy, in which "commanders must define the mission for the autonomous agent whether it be a human soldier or a robot."(Arkin, 2009, pages 37-38) Human and robot are equally embedded in this chain of command, with any failure to follow orders not revered as demonstrated moral virtue, but rather condemned as malfunction. Accordingly, to conceive of the human soldier as an autonomous agent in any non-contradictory sense requires a notion of autonomy inclusive of non-human killing machines, as well. Thus, we might reduce autonomous agency to "an embodied system designed to satisfy internal or external goals by its own actions while in continuous long term interaction with the environment in which it is situated."(Beer, 1995, page 173) But, this is just to say that an autonomous agent is simply a kind of efficient cause, that there is ultimately no distinction to be made between human and artificial agency, and that Beavers' fears of "ethical nihilism" have come true.

3. Rearranging the landscape

Interestingly, Kant also warned of the "quiet death" of morality, by the reduction of autonomy to "the physical order of nature."(Kant, 1780, page 7) But, before we review his defense of moral autonomy, it will pay to trace Ziemke's concept of this strong, "noumenal" autonomy typically reserved for humans to its origins in autopoiesis. "Autopoiesis," from the Greek meaning "self-producing," represents a rigid distinction between living and artificial agency according to which artificial agents are constitutionally incapable of autonomy, with an "allopoietic system like a robot deriving function from an external source," and the "primary function" of an autopoietic system "self-renewal through self-referential activity."(Amoroso, 2004, page 144) "Autopoiesis is the mechanism that imparts autonomy to the living,"(Luisi, 2003, page 52) with "the minimal form of autonomy" "a circular process of self-production where the cellular metabolism and the surface membrane it produces are the key terms."(Weber & Varela, 2002, page 115) So given, an autopoietic system is an organism that is both self-organizing, "one that continuously produces the components that specify it, while at the same time realizing it (the system) as a concrete unity in space and time, which makes the network of production of components possible,"(Varela, 1992, page 5) and far-from-equilibrium due to metabolic storage and "budgeting" of matter and energy in the development and maintenance of the "bodily fabric."(Boden, 1999, 2000)

We have already confronted some difficulties in distinctions based in the sources of goals, but there is much more to be said of "self-referential activity," a concept most important to the following section. According to an autopoietical account of agency, an organismThis bodily fabric emerges as a single, bound entity within "molecular space," with its properties (including semiological properties as signs and symbols are not abstract tokens but rather material tools) "structurally determined" by potential and actual chemical changes to the system.(Romesin, 2002) Co-emergent with the organism is the "niche," "the domain of interaction of the system with its surroundings, conditioning its possible ways of coupling with the environment,"(Rudrauf et.al., 2003, page 34) in terms of which it cognizes and acts in "selective coupling" with aspects of the environment, constituting the "operational closure" of the system., "the domain of interaction of the system with its surroundings, conditioning its possible ways of coupling with the environment."(Rudrauf et.al., 2003, page 34) This relationship dynamic between selective coupling and self-conditioning in a bubble of structurally determined significance constitutes the "operational closure" of the system, invitingleads to a view of cognition as "enacted," with the autopoietic system ultimately "creating its own world."(Luisi, page 58) So understood, an agent is a "self-producing coherence" bound to "maintain itself as a

distinct unity as long as its basic concatenation of processes is kept intact in the face of perturbations, and will disappear when confronted with perturbations that go beyond a certain viable range which depends on the specific system considered."(Varela, page 5)

Perturbations - "inputs" generally speaking - are responsible for two general classes of change, those within a "certain viable range" being "changes of state" through which the capacity of the system to self-organize, or adapt, is maintained, and "disintegrative changes" through which it is not.(Romesin) As "operational closure" extends from the molecular to cellular to organismic levels of organization, and upwards to social, cultural, and philosophical levels, an agent's niche can be understood as layers of increasingly conceptual order established in pro-active defense against disintegrative change, thus implying that "our minds are, literally, inseparable" not only from our bodies but from the environment as we experience it, thereby constituting a peculiar sort of "prison."(Rudrauf et.al., page 40).

As with the soldier cemented within a command structure, it is difficult to see how entrenchment within one's own "circular process of self-production" can ground autonomous agency in any non-contradictory sense. Effectively imprisoned in a semiological bubble of its own structural co-determination, this is as much as any artifact a portrait of *not*-autonomous agency. Comparatively, it seems that an agent with complete information about its embodied processes and origins, capable of specifying exact changes to its structure toward self-determined ends - swapping modules to suit particular purposes, as we might envision an artificial agent able to do - would enjoy greater autonomy than could any "living" thing. Thus, autopoiesis appears to be unnecessary for autonomous agency.

The case for an autopoietical foundation of autonomy is further weakened by the fact that the autopoietical distinction between living and non-living systems, and so the logic by which it "imparts autonomy to the living," is not very clear. For instance, Varela is reported to have not objected to the ascription of life to some synthetic molecular structures, Luisi's micelles, arguing that "our notion of life is heavily permeated by a religious bias (the notion of soul), which makes it difficult to freely use the word "life" for simple chemical systems," and that "Once one is liberated from these constraints, the term "life" may acquire a plainer and more usable meaning."(Luisi, page 58) However, in making this move, any necessary relationship between life and autonomy in any robust sense is severed.

In like spirit, one may argue for similar liberality regarding the term "autopoietic." Once liberated from constraints of cellular metabolism and surface membranes, autopoiesis can be fruitfully applied in the analysis of other systems including institutions and organizations (Goldspink and Kay, 2003, Hall and Dousala, 2010), legal systems (Vilaca, 2010) and social systems as a whole, (Leydesdorf, 1993) most famously through the work of Niklas Luhmann on whose account such systems are decidedly autonomous.(see Viskovatoff, 1999) Finally, with the autonomy of social systems, we are returned to the original, very plain and useful notion of autonomy with which this paper began.

In order to construct a comprehensive account of autonomy inclusive of both human and artificial agents while avoiding "ethical nihilism," however, we must review two further concepts from the autopoietical lexicon, "homeostasis" and "decoupling." A concept fruitfully developed by Antonio Damasio within the cognitive sciences, homeostasis (or better "homeodynamics") is the dynamic stability of a complex system achieved by balancing internal and external pressures through largely automated physical processes. On Damasio's account, as the cells of the body "gravitate" toward "fluid" states and away from "strained" "configurations of body state," they contribute to the "contents of feelings" as "both the positive and negative valence of feelings and their intensity are aligned with the overall ease or difficulty with which life events are proceeding."(Damasio, 2003, page 132) The positive association with objects that facilitate said stability transforms the world of objects into a space of value, such that "by the time we are old enough to write books, few if any objects in the world are emotionally neutral," with felt content rendered as "foundational images in the stream of mind" corresponding to "some structure of the body, in a particular state and set of circumstances."(pages 197 and 56)

Here, in the "gravitation" away from strained states, there is a basis for Wallach's "composite feeling" that is at the same time not limited to "living" systems. Consider, in this light, the molecule. The common representation of a molecule is that of a system sans strain, static and at rest. However, a more realistic image oscillates from strained configuration to strained configuration in dynamic equilibrium between forces internal and external. Now, a molecule doesn't "create its own world," but its presence does influence its environment, in special cases grounding the emergence of cellular and then organismic levels of organization, ultimately leading to evaluative content in the form of "the feeling of what happens," with even social systems emerging from "molecular space" by extension. This is not to say that "homeostasis" is the proper term for molecular dynamics. Rather, it is to say that everything in nature is a dynamic system, with homeostasis simply naming equilibrium seeking tendencies present in higher orders of organization. Following Alfred Kuhn (1974), we may suggest that all systems seek equilibrium in terms of their environments, and understand homeostasis as the general tendency for complex systems to compensate for forces of change while maintaining stability, integrity, and by extension even human dignity.

Indeed, it is this general tendency that ultimately grounds the emergence of autopoietical niches, themselves. Niches are spaces of cognition and action protective against forces of disintegrative change, fundamentally realized in Luisi's "living" micelles. A micelle insulates its interiority from potentially damaging external pressures, constituting a fundamental integrity "decoupled" from the environment, a proto-semiological bubble of self-production, effectively creating its own world within itself and of its own resources.

It is from this capacity to decouple from the environment, and not "life" however understood, that we can construct an account of strong autonomy equally inclusive of human and artificial agents. Following de Bruin and Kastner (2011), "decoupling" means "reducing direct effects of environmental stimulation and opening up possibilities for internally regulated behavior,"(page 10) thereby freeing an agent to act according to internal constraints rather than reflexively according to external triggers. These internal constraints extend throughout the range of agency, from chemical to symbolic, with capable agents creating their own purely conceptual worlds from their own cognitive resources. Decoupling thereby facilitates "hypothetical thought," a computationally demanding operation facilitated by formal constructs including counterfactuals and imperatives. "For example, hypothetical thought involves representing assumptions, and linguistic forms such as conditionals provide a medium for such representations."(Stanovish and Toplak, 2012, page 10)

Formal representations of hypotheticals further facilitate autonomy by representing situations potentially attainable through action and decoupled from an agent's structurally determined chemical-environmental entrenchment. This capacity to formally represent alternatives that guide action is "syntax autonomy." Syntax autonomy relies on "symbolic memory" through which agents gain "an element of dynamical incoherence with their environment (the strong sense of agency)."(Rocha, 1998, page 10) This formally mediated "incoherence" grounds the emergence of social and moral systems represented in theories of ethics and writs of history and law. Through these formal constructs, agents stipulate ends toward which they feel that actions should aim in a process "which involves the mutual orientation of agents in their respective cognitive domains to shared possibilities for future."(Beer, 2004, page 324) All told, this capacity to decouple from external pressures through symbolic mediation and to coordinate action to commonly beneficial ends over temporal limits far exceeding those of any constitutive agent is a powerful evolutionary force, known in traditional moral theory as "freewill."(see Juarrero, 2009)

4. Reinterpreting the tradition.

The preceding may seem to have strayed far from Kant's moral theory, when in fact we have merely plotted points for comparison in more recent discussions. For example, Kant anticipated the autopoietic distinction between life and artifact in terms of self-organization. In both, each part exists "by means of the other parts" as well as "for the sake of the others and the whole." However, in the natural organism "its parts are all organs reciprocally producing each other," so constituting "a whole by their own causality." Such an organized being is not a "mere machine, for that has merely moving power, but it possesses in itself a formative power of a self-propagating kind which it communicates to its materials though they have it not of themselves; it organizes them."(Kant, 1790, page 202) So understood, an organism is a "natural purpose" for Kant, "just the way we normally, prima facie and intuitively, view the living."(Weber and Varela, 2002, page 106)

However, also on Kant's account, far from autopoiesis imparting autonomy to the living, autonomy is hamstrung by self-productive requirements of the bodily fabric. "Life is the faculty a being has of acting according to laws of the faculty of desire."(Kant, 1788, footnote page 9) Meanwhile, autonomy, "autonomy of the will," or "freedom" as he variously calls it, "is a property of all rational beings," and to be free an agent must merely "regard itself as the author of its principles independent of foreign influences,"(Kant, 1785, pages 64-5) with such "foreign influences" including "the faculty of desire."

Accordingly, not only is life unnecessary for autonomy, it is a potential obstacle, calling into question the moral superiority presumed of human over artificial agents. Let's revisit the Kantian inspiration behind Ziemke's distinction between "noumenal" and "phenomenal" agency in this light. For Kant, when an agent conceives of itself as a "noumenon," he conceives of himself as a "thing in itself," "as pure intelligence in an existence not dependent on the condition of time," i.e. as if "immortal."(Kant, 1788 page 118) Here, we may understand "immortal" as free from the motivating necessities of embodiment, including the drives to maintain bodily integrity that so occupy the living agent, such that "he can contain a principle by which that causality acting according to laws of nature is determined, but which is itself free from all laws of nature."(page 118) So unfettered, an agent can focus on syntactic integrity, i.e. act in accord with the categorical imperative. This is why Kant equates autonomy of will with moral law.(see for example Kant, 1785, pages 62 and 66) "Autonomy of the will is that property of it by which it is a law to itself (independently of any property of the objects of volition)."(page 56) The difficulty for Ziemke's schema is that this independence is not necessarily observable, rendering, as we have already seen, any phenomenal ascription of autonomous agency suspect.

Digging deeper, there is in Kant a model of cognition and agency both accounting for these inner processes as well as giving us something to look for in ascribing autonomy. Excavating this model from traditional moral verbiage is difficult, however. To begin with, it is not enough for the Kantian moral agent to act freely toward just any ideal end, for instance out of purely scientific interest toward realizing the world as it is rather than as it appears, i.e. the "noumenal" rather than "phenomenal." Such agency is ultimately contingent on some "object of volition." Rather, on Kant's account, the ultimate promise of autonomous agency presents itself in the form of an "archetypal" world.(Kant, 1785, page 44) The archetypal world, variously referred to as the "kingdom of ends," the "summum bonum," the "supreme independent good," and even "God," is an ideal moral situation differing from the noumenal in that it is one with which an agent is "not in a merely contingent but in a universal and necessary connection," being the "destination" "assigned" by the moral law, "independent of animality," the "summum bonum" of the world.(Kant, 1788, page 165)

Kant's archetypal world appears to represent the mythical Christian "heaven," and such was in fact the model. However, Kant explicitly rejects the notion that recognition of any "God" is necessary for autonomy – and with this goes any requirement of a Christian "soul," for example.(Kant, 1788, page 133) Rather, Kant argues that an agent must merely hold three conceptions in order to be (potentially) autonomous: freedom (specifically, conceiving one's self as having the capacity to self-legislate, rather than serve bodily desires), immortality (conceiving one's self as if unbound by temporal constraints on the preceding), and God as the existence of a "supreme independent good."(page 137) "God" so understood is a destination, the archetypal end of action and "object of a will morally determined." Actions in accord with this ideal moral situation produce a deep moral pleasure subjectively realized as "harmony" with the extant realm consisting of all intelligent beings sharing in this ideal, with self-conception as "free" and "immortal" serving as limiting conditions on realizing this end.

Here, we are approaching an answer to Wallach's call for a comprehensive "platform for testing," noting that Kant also asks "What, then, is really pure morality, by which as a touchstone we must test the moral significance of every action?"(Kant, 1788, page 157) The key to answering this question lies in understanding how this "harmony" with the archetypal moral situation is possible for any "intelligent being," as it is this relationship that will finally bring Kant's model of moral cognition into the clear.

Intelligent being, synonymous with rational being, is the minimal condition for autonomy, the capacity to self-legislate. Autonomous action is determined by conceptions of law, rather than by "animality."(see Kant, 1788, pages 37 and 129) Thus, the Kantian portrait of agency is two-sided. One side is "immanent" through "transcendence," a "world of intelligence" and product of "intelligible being." The other is product of the immediate environment, animal attraction to "objects of volition" within the phenomenal world of sense.(page 108) These constitute two essential poles within the agent, one material and one ideal, as the Kantian agent "has two points of view from which he can regard himself, and recognize laws of the exercise of his faculties, and consequently of all his actions,"(Kant, 1785, page 70) and from which he may "pass judgment upon himself and his own actions."

As such, Kantian operational closure extends from the phenomenal world of appearances to the noumenal world of "things in themselves," understood as the archetypal world when one's own autonomous moral potential is fully realized. In conceiving of himself as free from material and temporal constraints, with an eye to the universal ideal situation the realization of which is his potential as an intelligent being, the agent "transfers himself in thought" "from the impulses of sensibility into an order of things wholly different from that of his desires in the field of sensibility," a situation in terms of which he does not imagine himself to be more comfortable, physically, but rather to have increased "intrinsic self-worth," a "better person" and "a member of the world of the understanding."(Kant, 1785, page 72) Accordingly, when we, as intelligent beings, "conceive ourselves as free, we transfer ourselves into the world of understanding as members of it and recognize the autonomy of the will with its consequence, morality."(page 70) "Moral pleasure" thus arises as an agent transcends embodied limitations and moves forward to the morally ideal "world of understanding" as his necessary and sufficient end of action.

Here, we find the ultimate bedrock of autonomous agency. The "fluid state" of one's self conceived as one's best possible self at once attuned to the best conceivable situation motivates the autonomous agent to realize that situation of its own freewill. The very possibility of morality arises in this realization, and Kant ties the survival of morality to its corresponding pleasure. Further, as the capacity to embody this condition is what gives autonomy to the autonomous, in this we have the terms to draw adequate distinction between degrees of agency, artificial or otherwise. Indeed, Kant writes that "No man is wholly destitute of moral feeling, for if he were totally unsusceptible of this sensation he would be morally dead."(Kant, 1780, page 30) Moreover, once this condition is realized, felt as a "good will," a moral agent is loathe to let it go, and regress into a relatively strained state. However, in order to understand why, we must review another concept from traditional moral theory, conscience.

5. Comprehending Autonomy

In Kant's words, conscience is "moral capacity" present as "an inward judge" "incorporated" into an autonomous agent's being from the position of the moral ideal, "as the subjective principle of a responsibility for one's deeds before God,"(Kant, 1780, page 41) i.e. from the perspective of the archetypal world. Conversely, to act contrary to the "dictates of conscience" produces a physical pain, "like grief, fear, and every other diseased condition," evidence of a proportional disharmony. So, even as moral pleasure reveals the possibility of morality, the self-disgust of inner discord reveals the possibility of immorality, providing a powerful motivation to morality, for "when a man dreads nothing more than to find himself, on self-examination, worthless and contemptible in his own eyes, then every good moral disposition can be grafted on it, because this is the best, nay, the only guard that can keep off from the mind the pressure of ignoble and corrupting motives."(Kant, 1788, page 163)

And, as conscience is of the fabric of rational agency, the moral duty to make "mankind in general his end" can be rewritten as "do nothing which by the nature of man might seduce him to that for which his conscience might hereafter torment him."(Kant, 1780, page 24) Thus conscience is the mechanism, or following Kant the "spring," of autonomous moral agency. Finally, so long as agency is conceived of as being bound by these two poles, good will and self-disgust, "ethical nihilism" is averted.

With this, we have in hand all of the necessary ingredients to answer Wallach's call for comprehensive ethics as "integrated being." The mechanism of this

integration is conscience. Conscience is nothing less than the mechanism of autonomous moral agency. It is the embodied capacity for the hypothetical comparison of one situation with others in terms with which the agent already cognizes and acts. It lays out possible ends of action as situations in which the agent should reach homeostasis, allowing for their comparison and relative evaluation, with the difference providing the motivation to move toward some rather than others. As the constitution of these hypotheticals proceeds from a limited sphere of individual experience, augmented by affective and effective mirroring as well as taught "top-down," the scope of conscience expands gradually over the course of operation. As terms increase, given sufficient resources, the agent may be able to balance greater numbers of dimensions, with associated dimensions bound together under single operators, simplifying the computational task. And, with the space of action mapped through this operation, conscience motivates the agent to seek situations with minimal strain between one's own and others' current and expected future situations, with the global minimum specified as the Kantian "summum bonum."

Some points of interest fall out of this portrait of autonomous moral cognition. For one thing, it naturalizes intension, understood as an internal, motivating and relatively evaluative felt strain, or tension, between conscientiously compared situations. It also naturalizes freewill, understood as embodied metabolic/energetic potential to construct and to act toward ends of one's own self-determination. These characterizations differ from those common to philosophy of mind, demanding accounts that cannot be fully developed here. However, they have been developed as aspects of the ACTWith model of moral cognition, and this model has been articulated in the contexts of model based reasoning and moral agency (White, 2010), psychopathy and moral psychology (White, 2012a), entropy and information ethics (White, 2012b) and autonomy in machine ethics (White, *in press*). For another thing, it is in terms of conscience that distinctions between degrees of autonomy can be consistently made. For example, Kant tells us directly that an agent would be merely "a marionette or automaton" without the tension between the sensible and the ideal made possible by conscience, with any sense of freedom a "mere delusion" deserving the name "only in a comparative sense, since, although the proximate determining causes are internal, yet the last and highest is found in a foreign land,"(Kant, 1788, page 102) i.e. not determined through conscience as a judge from the perspective of one's own projected moral perfection, but externally. Returning to the issues with which this paper began, this model of autonomous moral cognition answers Wallach's concerns about "integrated being," surprisingly enough by describing a being which integrates situations within the space of itself, in the process constituting the moral sentiment that is at once autonomy's signature. And, according to Beavers' proposal, it has been arrived at through some moral landscaping.

Through the preceding, it should be clear that most researchers in artificial agency go wrong in presuming that Kantian moral law must be pre-programmed as an explicit set of rules, when Kant takes great pains to show that the moral law co-emerges with the constitution of the rational agent. This constitution grounds autonomy, and with this fact the moral law emerges from a capacity to act regardless of material inclination, towards some universally good end, the guiding principle to which is formalized in Kant's categorical imperative. Accordingly, in response to Tonkins' (2009) "challenge to machine ethics," the real challenge in engineering fully autonomous AMAs lies in undoing prejudices stemming from misinterpretations of traditional ethical theory. The first step on this road to realize that these misunderstandings are only as temporary as are our personal commitments to them. Should autonomy be reduced to efficient causes, it is due only to our own lack of insight, our incapacity to free ourselves from our own embodied habits and conventions. So enslaved, so *not*-autonomous, it is no wonder that autonomy should forever remain a mystery.

References

[1] Amoroso, R.L. & Amoroso P.J. (2004) The Fundamental Limit and Origin of Complexity in Biological Systems: A New Model for the Origin of Life, in D.M. Dubois (ed.) *CP718, Computing Anticipatory Systems: CASYS03 - Sixth International Conference, Liege, Belgium,* August 11-16, 2003, New York: American Institute of Physics.

[2] Arbib, M.A. (2005). Beware the Passionate Robot, in Fellous, J.M., & Arbib, M.A. (ed.) *Who needs emotions? The brain meets the robot*. Oxford: Oxford University Press.

[3] Arkin, R. C. (2009). Governing lethal behavior in autonomous robots. Boca Raton: CRC Press.

[4] Beavers, A.F. (2012) Moral Machines and the Threat of Ethical Nihilism, appearing in Lin, P., Abney, K., & Bekey, G. A. *Robot ethics: The ethical and social implications of robotics*. Cambridge, Mass: MIT Press.

[5] Beer, R. (1995) A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72, 173-215.

[6] Beer, R. (2004). Autopoiesis and Cognition in the Game of Life. *Artificial Life*, 10, 3, 309-326.

[7] Boden, M. A. (1999). Is metabolism necessary? *British Journal for the Philosophy of Science*, 50, 2, 231.

[8] Boden, M.A. (2000). Autopoiesis and life. *Cognitive Science Quarterly*, 1, 117-145.

[9] de Bruin, L.C. & Kästner, L. (2011) Dynamic Embodied Cognition. *Phenomenology and the Cognitive Sciences.* Accessible at: http://www.springerlink.com/content/euxt674w7361j717/

[10] Damasio, A.R. (2003). *Looking for Spinoza: Joy, sorrow, and the feeling brain*. Orlando, FL: Harcourt.

[11] Goldspink, C., & Kay, R. (2003). Organizations as self-organizing and sustaining systems: a complex and autopoietic systems perspective. *International Journal of General Systems*, 32, 5, 459-474.

[12] Hall, W.P., and Dousala, S. (2010) Autopoiesis and Knowledge in Self-Sustaining Organizational Systems. *4th International Multi-Conference On Society, Cybernetics And Informatics*. Orlando, Florida, USA.

[13] Juarrero, A. (2009) Top-Down Causation and Autonomy in Complex Systems. In Murphy, N.C., Ellis, G.F.R., & O'Connor, T. *Downward causation and the neurobiology of free will*. Berlin: Springer.

[14] Kant, I. (1780) *The Metaphysical Elements of Ethics*. trans. Abbott, T.K. Pennsylvania State University Electronic Classics Series (2005). http://www2.hn.psu.edu/faculty/jmanis/kant/metaphysical-ethics.pdf

[15] Kant, I. (1785) *Fundamental Principles of the Metaphysic of Morals*. trans. Abbott, T.K. Pennsylvania State University Electronic Classics Series(2010). http://www2.hn.psu.edu/faculty/jmanis/kant/Metaphysic-Morals.pdf

[16] Kant, I. (1788) *The Critique of Practical Reason*, trans. Abbott, T.K. Pennsylvania State University Electronic Classics Series (2010).

http://www2.hn.psu.edu/faculty/jmanis/kant/Critique-Practical-Reason.pdf
[17] Kant, I.,(1790/1914) Bernard, J.H., & Liberty Fund. *Kant's Critique of Judgement*. London: Macmillan. http://files.libertyfund.org/files/1217/Kant_0318_EBk_v6.0.pdf
[18] Kant, I. (1796). *The Metaphysics of ethics*. Edinburgh: T. & T. Clark. Available at: http://oll.libertyfund.org/title/1443 on 2012-01-24
[19] Kuhn, A. (1974). *The Logic of Social Systems.* San Francisco, CA: Jossey-Bass.
[20] Leydesdorff, L. (1997) Is society a self-organizing system? *Journal of Social and Evolutionary Systems*, 16,3, 331-349.
[21] Luisi, P. L. (2003). Autopoiesis: a review and a reappraisal. *Die Naturwissenschaften*, 90, 2, 49-59.
[22] Moor, J.H. (2006) The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21, 18–21.
[23] Moor, J.H. (2007) Taking the Intentional Stance Toward Robot Ethics. *APA Newsletter*, 6, 14-17.
[24] Reath, A. (2006). *Agency and autonomy in Kant's moral theory*. Oxford: Clarendon Press.
[25] Rocha, L.M. (1998). *Syntactic autonomy*. Los Alamos National Laboratory. Washington, D.C: United States. Dept. of Energy.
[26] Romesin, H. M. (2002). Autopoiesis, Structural Coupling and Cognition. *Cybernetics and Human Knowing*, 9, 5-34.
[27] Rose, S. (1998). *Lifelines: Biology beyond determinism*. Oxford: Oxford University Press.
[28] Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J. P., & Le, V. Q. M. (2003). From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. *Biological Research*, 36, 1, 27-65.
[29] Smithers, Tim (1997) Autonomy in Robots and Other Agents. *Brain and Cognition* 34: 88-106.
[30] Stanovich, K. & Toplak, M. (2012) Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*. http://www.springerlink.com/content/x461x01027625w35/
[31] Tonkens, R. A challenge for machine ethics. *Minds & Machines*, 19, 421–438, 2009.
[32] Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12, 3, 243-250.
[33] White, J.B. (2010). Understanding and augmenting human morality: An introduction to the ACTWith model of conscience. *Studies in Computational Intelligence*, 314: 607-621.
[34] White, J.B. (2012a) An Information Processing Model of Psychopathy and Anti- Social Personality Disorders Integrating Neural and Psychological Accounts Towards the Assay of Social Implications of Psychopathic Agents. In *Psychology of Morality*. New York: Nova Publications.
[35] White, J.B. (2012b). Infosphere to Ethosphere: Moral Mediators in the Nonviolent Transformation of Self and World. *International Journal of Technoethics*, 2, 4, 53-70.
[36] White, J.B. (book chapter, in press) Manufacturing Morality: A general theory of moral agencygrounding computational implementations: the ACTWith model. In *Computational Intelligence*. New York: Nova Publications.
[37] Varela, F. (1992). Autopoiesis and a biology of intentionality. Appearing in *Proceedings of a workshop on Autopoiesis and Perception*, 4–14. ftp://ftp.eeng.dcu.ie/pub/alife/bmcm9401/varela.pdf
[38] Vilaca, G.V. (2010) From Hayek's Spontaneous Orders to Luhmann's Autopoietic Systems. *Studies in Emergent Order*, 3, 50-81.
[39] Viskovatoff, A. (1999). Foundations of Niklas Luhmann's Theory of Social Systems. *Philosophy of the Social Sciences*, 29, 4, 481-516.
[40] Weber, A., & Varela, F.J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and Cognitive Sciences, 1*, 2, 97-125.
[41] Ziemke, T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems*, 91, 401-408.