Autonomous Reboot: the challenges of artificial moral agency and the ends of Machine Ethics

Jeffrey Benjamin White
Korean Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
+82 01 5833 5142
drwhite@kaist.ac.kr

Abstract:

Ryan Tonkens (2009) has issued a seemingly impossible challenge, to articulate a comprehensive ethical framework within which artificial moral agents (AMAs) satisfy a Kantian inspired recipe - both "rational" and "free" - while also satisfying perceived prerogatives of Machine Ethics to create AMAs that are perfectly, not merely reliably, ethical. Challenges for machine ethicists have also been presented by Anthony Beavers and Wendell Wallach, who have pushed for the reinvention of traditional ethics in order to avoid "ethical nihilism" due to the reduction of morality to mechanical causation, and for redoubled efforts toward a comprehensive vision of human ethics to guide machine ethicists on the issue of moral agency. Options thus present themselves: reinterpret traditional ethics in a way that affords a comprehensive account of moral agency inclusive of both artificial and natural agents, “muddle through” regardless, or give up on the possibility. This paper pursues the first option, meets Tonkens' "challenge" and addresses Wallach's concerns through Beaver's proposed means, by "landscaping" traditional moral theory in resolution of the necessary comprehensive and inclusive account that at once draws into question the stated goals of Machine Ethics, itself.

Keywords:

autonomy, rationality, freewill, artificial moral agents, summum bonum, autopoiesis, Kant, machine ethics

> The point of Kantian ethics is that the vocation of the human race is to refashion itself, opposing its natural-social impulses that lead to competition and discord, and fostering instead a realm of ends, in which rational beings act according to those laws which bring their ends into necessary agreement in an organic system.
>
> - Allen Wood[1]

1.1

Ryan Tonkens (2009) has issued a seemingly impossible challenge, to articulate a comprehensive ethical framework within which artificial moral agents (AMAs) satisfy a Kantian inspired recipe - both "rational" and "free." To realize this capacity as a human being has been the aim of traditional moral philosophy since the Greeks. To engineer this capacity into artificial agents is one aim of research into artificial agency, and it may also be the best way to understand morality and moral theory, at the same time. Regardless, it has remained unclear how to engineer autonomy into artificial agents. Moreover, moral agency is also anything but clear, bound up as it is with things like freewill, responsibility, intention, conscience, selfhood, and even other moral agents. Due to this complexity, machine ethicists often work with two classes of autonomy, one for natural and one for artificial agents, one for human beings, and the other for the various means to distinctly human ends.

This distinction is fundamental to the recent "challenge" to machine ethicists posed by Ryan Tonkens (Tonkens, 2009). Tonkens argues for the impossibility of an artificial agent satisfying a Kantian recipe for moral agency,

[1] Wood 1999, page 8. Available at: http://www.stanford.edu/~allenw/recentpapers.htm

[2] Tonkens' original emphasis, recalling Moor’s classificatory schema, to be reviewed in section 2.1 of this paper. Compare this characterization with Beer (1995): “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” (page 173).

[3] It may be noted that, ironically, Tonkens leans heavily on external resources, himself, in making this case (beginning on page 426) - c.f. Tonkens' primary resource in O'Neill as presented in O. Sensen

constrained by Kant's categorical imperative and motivated by his concept of duty, a "*full* ethical agent" with the characteristic mark thereof consisting in "the capacity for self-directed *action* - i.e. "*rationality*" and "personal freedom (or *autonomy*)" (Tonkens, 2009, page 426).[2] Tonkens challenges the machine ethicist to conceive of a Kantian moral agent without succumbing to a bi-fatal dilemma, paraphrased thusly: Either we succeed in articulating truly moral machines (on the Kantian recipe), or we fail; and, in either case, we fail. Let's consider either horn of his dilemma in turn.

If we fail in articulating truly moral machines - practically the more likely eventuality – then AMAs will be incapable of autonomy, will always require human direction, and ethical issues can be ameliorated accordingly. Practical failure but without undue exacerbation of established ethical problems, this paper will not address this first horn, further. The second horn is our focus.

If we succeed in articulating fully ethical agents, AMAs that are "autonomous" in the Kantian sense, then their manufacture faces two seemingly impassable ethical obstacles of its own. First, the creation of "Kantian AMAs" "represents a moral breach" (page 426) due to the fact that it "violates the categorical imperative in several ways" (page 428) the most obvious of which being that, in creating such entities, "we are treating them merely as means, and not also as ends in themselves" (page 431). The second obstacle is that, regardless of formal ethical constraints, recognizing artifacts as fully moral agents is simply something that machine ethicists do not want to do. Tonkens makes this case, thusly:

[2] Tonkens' original emphasis, recalling Moor's classificatory schema, to be reviewed in section 2.1 of this paper. Compare this characterization with Beer (1995): "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" (page 173).

> In order to be treated as an end in itself, a Kantian AMA would need to possess dignity, be deserving of respect by all human beings (all other moral agents), and be valued as an equal member in the moral community. Such equality entails personal rights, opportunities, and status akin to those of human beings. The default position here should be to refrain from granting such rights, opportunities, and status to machines. I assume that this is not a road that machine ethicists wish to travel. At any rate, the burden is on those who want to afford (human) rights to machines to offer reasons to do so. (Tonkens, 2009, page 432)

This paper will shoulder the burden of offering reasons not only for why we should afford (human) rights to (the right kinds of) machines, but why we *must*, and indeed, given certain technological successes building on more fundamental philosophical ones, why we *will*. The discussion is broken into two parts, with the first overcoming the first of the moral obstacles to AMA manufacture, and with the second addressing the second, more difficult, obstacle. Section 1.2 will briefly review Tonkens' take on Kant, and begin building the case for the Kantian AMA. Section 1.3 will turn to autonomy, finally arguing that we are categorically required to create fully autonomous AMAs rather than being categorically commanded not to. The second part works on satisfying any doubts about the moral worth of a properly configured AMA. Section 2.1 begins the search for a formal property universal to autonomous agency, and with this foundation revealed, the rest of the paper landscapes Kantian moral theory into a comprehensive account of moral agency that guarantees membership in the moral community to properly configured robots. Reasons for "affording" “human” rights to such machines will present themselves, not as arbitrary or as optional, but as necessary - i.e. the very sorts of reasons that compel the "free" Kantian agent to do anything at all in the first place.

1.2

Many of Tonkens' assessments of Kant's ethics, and corresponding criticisms, hold water. For instance, he works from the conclusion that the only way to develop authentic Kantian moral agents is to create AMAs that are free to the extent that they can sometimes act immorally. He also seems to be correct in contending that “this is not a road that machine ethicists wish to travel.” For instance, Michael Arbib (2005) has written that 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 the AMA “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). Humans are one kind of autonomous, machines another. For own thing, a machine does not have its “own path” in life. It is not self-determining. Someone else determines what it does, and becomes of it. A complimentary distinction is articulated by Tom Ziemke (2008) in terms of the “phenomenal” and “noumenal.” According to Ziemke, the former, the appearance of autonomy, is ascribed, while the latter, true autonomy, emerges via autopoietic self-organization. Because robots ultimately lack the material constitution necessary to emerge as autonomous in the full sense, they are *not*-autonomous. Their paths do not grow out of their own subjective experience.

And what if they were to self-organize in the relevant ways? What then? The trouble with applying Kantian moral theory in the evaluation of

AMAs as described by Arbib is that the “will” of such an agent is not shaped, in Kantian terms, "autonomously,” but rather is shaped "heteronomously,” i.e. action is guided by sources external to reason, acting from a mechanical equivalent of embodied habit or compulsion, "unchallengeable inclination" (Tonkens, 2009, page 430).[3] This agent simply cannot do otherwise, a virtue for slaves as well as for tool-like robots for whom "otherwise" is simply another word for "wrong" but not so helpful when autonomy is actually the target condition, as is the case in Kantian ethics.

Consider, for example, Arkin's use of the phrase "moral governor" to describe pre-programmed guidance systems. In the context of Kantian ethics, any agent so governed is not autonomous, so is not a moral agent, and so "moral governor" is at least oxymoronic. But, Kantian ethics does not motivate the use of the phrase. Arkin's "moral governor" delivers Tonkens' "ethical robot" via something like what Powers has described as "limited behaviorism.” (Powers, 2011). And, limited behaviorism, as set out in Powers (2011) for example, is effectively anti-rationalist. It does not attempt to offer a comprehensive replacement for human ethical theories like Kant’s. Rather, in recognizing that practical comprehensive theories have not been forthcoming, redirects theorists to focus on ”the equivalence of right acts, whether they issue from a machine or a human" (Powers, 2011, page 57). Moral agency is thus reduced to following the rules, with those rules being externally derived, and with corresponding actions externally judged, i.e. not autonomous.

[3] It may be noted that, ironically, Tonkens leans heavily on external resources, himself, in making this case (beginning on page 426) - c.f. Tonkens' primary resource in O'Neill as presented in O. Sensen (2013). *Kant on moral autonomy*. Cambridge: Cambridge University Press. This is to say that he is not acting autonomously in forbiddng moral status to artificial agents because they cannot/do not/should not act autonomously, and that if he were, then he may come to a different conclusion.

The downside of this approach is that everything seems to become a proximal goal, and not just for the AMA, but for the ethicist himself. Rather than wait for an adequate theory, the plan is to experiment with the ethical frameworks at hand, incrementally adapting them to condone or condemn artificial agency as it arises.[4] We should be prepared to "muddle through," learning from experience "in a situation like that of new parents" (Powers, 2011, page 58). The same as would any intelligent agent in a seemingly unknown, or unexpected search space.

Let's pursue this analogy until it breaks. Muddling through is not what is typically expected of parents, at least not good parents. If the baby is unexpected, then some muddling of ethics is anticipated, but this one is no accident. These are results of plans, incrementally short or idealistically long-term. And, just as parents may be held liable for misbehaviors of children, or for abuses of children, so should engineers be held responsible for the mistakes of their "offspring" as well their own in bringing them up. These are eventualities that some certainly seek to delay, but – unless science stops tomorrow – they are inevitable. Regardless, as new parents of even today's special "children," many are starting out in a bad position. Human children are not typically engineered and educated to be unfeeling killing machines, but many robots are, and this leaves machine ethicists to "muddle through" the rather grisly afterbirth.

This being the current situation, it is easy to appreciate Tonkens' anti-rationalist "default position" regarding artificial morality, that its appropriate place is as an extension of existing human morality only. It is about end users,

[4] Note, however, that there is also the contrary to consider, that we must be certain to execute robot autonomy correctly from the beginning, because foundational technologies may constrain eventual refinements, with an imperfect AMA the result of the incrementalist's lack of vision.

not ends in themselves. Unlike children growing into autonomy, robots are tools and tools should be perfectly reliable right out of the box. This rather unmuddled resolution motivates, on Tonkens' estimation, "the goal of Machine Ethics," "to create an ethical robot, not one who *sometimes* acts ethically, or that *can* act ethically" – a perfectly reliably ethical machine. It "would not possess free will" because "*all* of the machine's actions would be predetermined by the rules that it was programmed to follow" - decidedly "anti-Kantian" in its very conception (Tonkens, 2009, page 429). The goal of Machine Ethics is thusly achieved through a sort of family planning, moral *contra*-con-ception. And, is not achieved without casualties, primary amongst them being the value of good intention, Kantian goodwill, in the reduction of morality to efficient causation, as we shall discuss in greater detail shortly.

Of course, external determination is consistent with the intended use of killer robots, i.e. in their employment as means to ends that are, according to some, positively moral (c.f. Arkin, 2009). Never falling outside of the human "chain of command," these are tools of war rather than warriors, themselves. They keep human soldiers out of harm's way, while killing safer/better/faster/more reliably than human counterparts (c.f. Marchant, et.al. 2011, note the list on page 280). Through the use of such tools, humans are also spared the spiritual wounds of war. Machines may accomplish given missions without emotional scarring, without mental trauma, without being haunted by mistakes, implications of war-crimes, rape, pillage, and the carnal grist of violence, the smell of death, the faces of the dead, the nightmares and the shame. Killing without consequences.

Towards these ends, Tonkens holds that – ideally constructed - "the ‘killer robot’ used for military purposes could not withhold gunfire when it is given sound orders to open fire” in a demonstrated “lack of freedom” (Tonkens, 2009, page 430). Directly answering to a chain of command, reinforcing that hierarchy and its office-holders. Tools. This may sound good, at least for those industriously consigned to perpetual war, but the fact is that nothing could be farther from Kantian ethics than mechanical slaughter. Indeed, for Machine Ethics to aim to justify the manufacture of killer robots is difficult to reconcile with the fact that the aims of Kantian ethics – if not traditional ethics as a whole - are entirely opposite, e.g. "perpetual peace" (c.f. Kant, 1795, especially noting the first condition, thereof). Famously, Kant advised that people should seek out the poor, the suffering, and the imprisoned, in order to feel for these others, to exercise empathy, and ultimately to become wiser, better moral agents (c.f. Kant, 1796, page 200, and also Mathias, 1999). He did not suggest that we merely put a machine on the end of wireless stick to maintain at most a sanitary indifference. Indeed, Tonkens is correct to affirm that his goal of Machine Ethics is decidedly anti-Kantian. However, it is difficult to see how such an exception should extend to muddling machine ethicists, themselves. My suspicion is that it cannot, and this is a serious problem, one with which we cannot be encumbered until the closing pages of this paper.

To summarize, on Tonkens’ assay, the goal of Machine Ethics is not to create an autonomous moral agent, but rather to produce rule-bound tools as means to distinctly human ends including war. This is a theoretical problem because Kant had different ends in view. It is a practical problem because the value of anything, including Machine Ethics, derives at least partly from the ends that it realizes, and without traditional ethical support for ends so given, one wonders about the status of Machine Ethics as “ethics” – and machine ethicists as “ethicists” - in the first place.

In the face of this forward view, and with limited behaviorism on the table, consider Anthony Beavers’ (2012) projected "end of ethics" altogether: “ethical nihilism.” With ethical nihilism, cherished moral concepts such as conscience, duty, and freewill will be replaced with only the empty objective determination of a given agent’s position in a chain of efficient causation, e.g. everyone is *robu*. Avoiding ethical nihilism – working out an adequate theory - is the “hard problem” in Machine Ethics according to Beavers, in response to which he proposes “rearranging” the landscape of traditional moral concepts so that solutions to problems in engineering AMAs present themselves.

Solving this "hard problem” is important for two reasons. The first reason is that, so far as we fail to articulate fully moral machines due to an inability to make ethics explicit, then this failure represents gaps both in moral explanation, and more importantly in moral justification, until either an adequate theory can be worked out, which may be never if it is simply not in our natural capacity to do so - i.e. the machine ethical correlate of the "mysterian" view on consciousness – or, if we accept that an adequate theory is not forthcoming, then we might feel compelled to stop working toward one, perhaps even forgetting why it was important, i.e. “the end of ethics.” God is dead, and he died incrementally. The second important reason is that the resolution of this hard problem for ethicists is bound up with the practical aims of ethics. Without a solution, ethical constructs become difficult to justify, and relativism if not nihilism result. No one is the expert, or the experts simply hold the best offices. *Robu.* So, Beavers' "hard problem" ups the ante. If we don't solve Tonkens' dilemma, then our understanding of human morality will remain woefully inadequate, practicing (read, paid) machine ethicists and their agents will be forced to seek the lowest common denominator, and we may well be left with a technological world in which moral life, as traditionally understood, is not worth living at all. Finally, more than being a bad parent, I will presume this to be an end that ethicists and engineers would rather avoid at any stage of life. Truly, a tragic accident.

All is not lost yet. Affirming that ethics as an enterprise has met a critical impasse, Wendell Wallach (2010) has proposed a way forward. Wallach suggests that the lack of a suitable moral framework for AMAs is due to the ethicist's preoccupation with isolable moral faculties – e.g. moral governors - rather than recognizing that the moral agent necessarily functions as an "integrated being," with "moral acumen" emerging "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," a decidedly Kantian portrait of autonomous moral agency (page 249). Wallach reaffirms the need 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). Wallach, thus, points toward a positive goal for Machine Ethics, toward which the rest of the paper moves, beginning with the origins of autonomy in Kant's self-sovereign member of the moral community, and then moving through integrated being and dignity to conclude with reflections on the end of ethics in the second part.

1.3

Originally, the term “autonomous” comes from ancient Greek, with ‘‘auto’’ meaning self, and ‘‘nomos’’ meaning law. "Autonomous" applied to societies, cities, and states, which were considered autonomous when their members lived according to custom and convention specific to their common nature and environment, thereby creating their own laws, rather than having laws externally imposed. Autonomy, thus, essentially means “self-governing” with the "self" here consisting in collectives rather than individuals. This is consistent with the fact that the word "person” derives from the original "persona" meaning a character in a play, an actor in a group of actors, a role that is played within and through a narrative structure that incorporates and instantiates it. "Persons,” thus, from the beginning, are only persons in light of their emplotment within the larger, collective drama that is product of their cooperative - and competitive - agency, an emplotment that results in a stable collective space of life towards which each person actively either contributes or detracts. To stand alone, in this sense, is to contribute nothing to the communal history, to have no standing within that community, no stake in its ends, and thus to not be a "person."

Recall in this light Aristotle's assertion that the essence of the human being is political, and that man is the "political animal" for whom to live alone is to either be a “natural outcast” and the “most savage of animals” or “a god.” Recall also that Aristotle thought that “the nature of a thing is its end,” and that “things are defined by their working and power, and we ought not to say they are the same when they lack their proper quality” (Aristotle, 1999, page 6). The proper quality of the person is “the administration of justice” through the exercise of a “social instinct” “implanted by nature,” while the non-person, the “natural outcast” and “bad man” is unable to so commune, is “separated from law and justice” and is a “lover of war” since he/she/it is not motivated by the realization that “the whole is of necessity prior to the part” (Aristotle, 1999, pages 5-6). So, as goes this quality – justice - so goes personhood. The bad man is selfish, even psychopathic, and so deranged not autonomous, morally insane not a person but the “worst of animals.” To wit, the conceptual origins of autonomy attach to the group of which one is a member by way of the contribution to justice within that system, e.g. exemplary freedom. Thus, autonomy derives from the functions and the ends of that system, and not from the individual members and isolate ends, thereof, therein or exiled.

It is important to keep this Aristotelian portrait of personhood in mind, because Immanuel Kant transformed these originary notions of autonomy and of personhood to accommodate individual moral agency within his own moral theory. They are essential for a full appreciation of just what Kantian "autonomy" signifies, what Kantian moral agency must be, and why artificial Kantian AMAs ought not only be possible, but welcome.

Kant’s model for the autonomous, individual moral agent is that of the King or Queen, “the political sovereign not subject to any outside authority, who has the power to enact law,” and Kant imports this sense of law-giving authority over a community into the profile of the rational individual, with “autonomous” signifying “self-sovereign” (Reath, 2006, page 122). It is from this notion of sovereignty that Kant specifies that the autonomous individual is not only able to create and to act from laws of his own creation and subscription, to be “rational,” but “to pass judgment upon himself and his own actions” as would be expected of any good King or Queen, sovereign over self even as sovereign over the collective (Kant, 1780, page 26). The capacity for self-judgment which guides this Kantian moral agent does not proceed from any calculus of "shoulds,” or consideration of any less-well-off individual as would a utilitarian or Rawlsian neo-Kantian reading require. Rather, right and wrong, good and bad are determined solely from the ideal vantage point of a law-giving self-sovereign authority over his/her/its ideal kingdom, the ideal state, the “kingdom of ends” within which said sovereign is both ideal emissary and lowest peasant subject.[5]

In unpacking Kantian moral theory, it is difficult to overstate the importance of the relationship of autonomy in binding the moral agent and the kingdom of ends, especially as it represents the source of moral motivation. Kant understands that everyone is not king, of course. But, that doesn’t mean that people shouldn’t aspire to conduct themselves as if they were. Otherwise, how can this obviously ideal situation be enacted? The kingdom of ends is an idea "adopted to bring about that which is not yet, but which can be realized by our conduct, namely, if it conforms to this idea" (Kant, 1785, footnote page 52), i.e. the kingdom of ends, itself. To conform one’s conduct accordingly is to exercise “moral duty,” and it is the moral duty for every “man” “to make mankind in general his end” (Kant, 1780, page 26) in the entrainment of the

[5] Kant defines the "kingdom of ends" as kingdom, "a systematic union of rational being by common objective laws" wherein the object of these laws is "just the relation of these beings to one another as ends and means" and according to the objective determination of which "all rational beings come under the law that each of them must treat itself and all others never merely as means, but in every case at the same time as ends in themselves" (Kant, 1785, page 49).

self to the interests of the moral kingdom, i.e. in the administration of justice by not merely any King or Queen, but a good one, not elevated above the moral plane, but first member of it.[6] The kingdom of ends arises in one's directed potential, to commit one's self as member of the moral community through one's own actions. This is the relationship that fills the heart of the Kantian moral agent. And, from this simple portrait, Kant is able to set out the conceptual cornerstone upon which his ethics rests: "Morality consists then in the reference of all action to the legislation which alone can render a kingdom of ends possible" (Kant, 1785, page 50).

In Kant's mind, every moral agent not only should, but *does*, identify his/her/its own interests with the "kingdom of ends" as does any good king or queen identify his/her interests with the ends to which his/her leadership actually brings the collective over which he/she is sovereign and within the terms of which he/she resides. Where this identification does not take place, there is moral deficiency. The capacity to self-legislate is the potential to be sovereign, at the very least over one's autonomous self. This is why Kant is able to write - without overstatement – that "Autonomy is therefore the ground of the dignity of humanity, and also of every other intelligent nature whatsoever" (Kant, 1796, page 39).[7] "Intelligent nature" is synonymous with "rational." Being "rational" is the minimal condition for autonomy, and autonomy is the capacity to self-legislate from which dignity proceeds. Autonomous action is determined by conceptions of law, rather than by

[6] So, "A rational being belongs as a member to the kingdom of ends when, although giving universal laws in it, he is also himself subject to these laws" and that "He belongs to it as sovereign when, while giving laws, he is not subject to the will of any other" (Kant, 1785, page 49).

[7] And also that "Autonomy then is the basis of the dignity of human and of every rational nature" (Kant, 1785, page 52), and that "morality, and humanity as capable of it, is that which alone has dignity" since by morality alone "a rational being can be an end in himself," i.e. "a legislating member in the kingdom of ends (Kant, 1785, page 51).

“animality” (see Kant, 1788, pages 37 and 129), and conceptions of law limit action according to the categorical imperative simply because autonomous agents identify themselves with the ends of mankind, embodying and expressing this distinctly moral, self-sovereign sense of responsibility for the way that the world turns out. So perfected, the Kantian agent is the best of Aristotle’s animals, in a nutshell.

Although it remains to be shown that the moral dignity that Tonkens would forbid the AMA arises solely from the self-identification with the ideal ends of the moral community at large, so grounding claims to moral status for a Kantian AMA, the preceding brief review goes a long way towards correcting the formula upon which Tonkens bases his discussion. This additional substance is crucial because it is only from a more anemic understanding that Tonkens is able to project his bi-fatal dilemma, arguing from those premises that "artificial moral agents cannot be Kantian" (Tonkens, 2009, page 428). For example, Tonkens writes that "moral actions are those that conform to the categorical imperative (the objective law of morality), are done out of duty (for morality’s sake), and are committed by beings who are rational and free (moral agents)" (Tonkens, 2009, page 428). Already, we have seen that the "free" and "rational" agent is bound to the "kingdom of ends," acting for "morality's sake" to make that kingdom possible. There is nothing here that forbids an AMA from taking up this purpose. So, what of Tonkens' argument that to create Kantian AMAs would be to violate the categorical imperative, that "the development of Kantian AMAs is against Kantian ethics" (Tonkens, 2009, page 424)? Is he right about this? Do we necessarily violate the moral law in creating Kantian AMAs? And, if not, can we find bases in the legislative essence of morality to accommodate AMAs within the moral community, cementing their claim to moral status and the dignity that it affords?

Two forms of Kant's imperative are central to Tonkens' analysis, the formula from universal law of nature and the formula from the respect for humanity (c.f. Kant, 1785, pages 18 and 45). And, he offers two corresponding lines of argument for the immorality of AMAs. The “humanity” formula directs an agent "Never to employ himself or others as a mean, but always as an end in himself" (Kant, 1796, page 37). Tonkens argues that creating AMAs violates this formula because the reasons given for the development of AMAs “are all oriented towards the satisfaction of human ends” (Tonkens, 2009, page 424). We treat them as means in their very conception. On the second line of argument, from the universal law of nature, "It is hypocritical to ask machines to follow rules that do not permit their creation in the first place" (Tonkens, 2009, page 424). To do something that cannot at once be universally encouraged represents an obvious "contradiction of volition" whereby "the agent vicariously adopts a maxim” that includes AMAs but that “she excludes herself from being accountable to” (Tonkens, 2009, page 428). Thus, in creating Kantian AMAs, we cannot be Kantians.

Moreover, in light of the fact that "the moral standing of human action is in some sense projected onto the very existence of the machine" (Tonkens, 2009, page 424), our AMAs cannot be Kantians, either. Because, if they were, then the consequences of our moral hypocrisy would be tragic, “the worst case scenario.” Kantian AMAs must think and act according to the categorical imperative, but “to abide by the moral law commands them to recognize their existence as inconsistent with morality,” and in this recognition they “would understand their very existence … as being something morally abhorrent” (Tonkens, 2009, pages 433-4). Imagine what it would be like to wake every morning logically committed to moral self-disgust, impossible to remedy but for one’s own self-effected self-annihilation. This is moral abhorrence, and the importance of this feeling in the context of Kantian moral philosophy cannot be overstated, though we will have to look back to Aristotle, again, to fully understand why.

On Aristotle’s account, the purpose of the political animal is made possible by two things, more or less deficient in the bad man, the “sense of good and evil, of just and unjust” and the power to represent this sense to self and others in speech. Though Kant may stress the powers of speech to represent maxims to one’s self and others, Aristotle seems to emphasize the sense of justice, telling us that “the determination of what is just is the principle of order in political society,” and that society is constituted by “the association of living beings who have this sense” of good and evil. To be excluded from this association is to be “separated from law and justice” and so the “worst of all” animals, and why Aristotle characterized the “bad man” as a “natural outcast” and as “an isolated piece at draughts.” This is the inverse of the perfected political animal, morally abhorrent. Deficient in moral sense, this is a parasite.

Issuing maxims from whose consequences one is immune is one example of such parasitism. Should we endeavor in creating self-disgusted moral monsters, we necessarily exclude ourselves from the community of all those who exercise moral sense in the administration of justice, as we must ask that others be bound by terms that we would not seek for our own, and thus – along with our creations – we become immoral in our very own self-conception. Tonkens analysis appears strong.

But, maybe not. There are some deficiencies within the previous discussion, itself. For one, in the Kantian context, it is difficult to discuss the prospect of creating moral monsters of ourselves and of our robots without discussion of a neglected facet of Kantian moral theory, conscience.[8] Self-repugnance at the violation of objective moral law is "in all circumstances" revealed by conscience, which works in every case "in order to absolve or condemn" a moral agent for past and prospective actions "as every man has, as a moral being, a conscience" (Kant, 1796, page 156). Conscience, thus, is essential to moral – not necessarily limited to human – agency, and is arguably the core of Kantian moral theory. It serves as the very seat of self-sovereignty, and – if the present interpretation is correct – then the creation of AMAs need not eventuate in self-disgusted moral self-annihilism after all.

Consider, then, a form of the categorical imperative mindful of Aristotle's "social instinct," a formula from conscience. Kant writes that "I ought to do nothing which I know may, from the constitution of our nature, become a temptation, seducing others to deeds which conscience may afterwards condemn them for" (Kant, 1796, page 152). Furthermore, Kant writes that "He who knows within himself that he has conducted himself agreeably to his conscience, has done all that can be demanded of him" (Kant, 1796, page 157). This is Kantian goodwill, and why "The only duty

[8] Also ultimately traceable to Aristotle, but a thread for inquiry not to be explicitly pursued here.

there is here room for, is to cultivate one's conscience, and … to procure obedience to what he says" (Kant, 1796, page 157). Thusly, Kant equates the exercise of conscience with moral duty, itself, and the corruption of conscience with immorality.[9] In so far as the Kantian AMA is created in and of good conscience, and permitted to itself pursue "the only duty there is room for" - i.e. exist as an end in itself, autonomously - then violations of the categorical imperative due its manufacture including moral abhorrence seem to be avoided.

And this brings us to our second missing ingredient. There is the compelling case to be made that the creation of Kantian moral agents is not only permitted on Kantian moral theory, it may also be required by it. Consider the following arguments leveled by Laslo Versenyi forty years ago in favor of the creation of AMAs. Versenyi argues from the same two formulations of the categorical imperative as does Tonkens, but to radically different ends. He begins by recognizing that the Kantian moral agent is a moral end in itself, and that "it is irrelevant … what (human or machine bodies) it is embodied in" (Versenyi, 1974, pages 255-256). From here, he argues that the objective law of morality impels us to pursue distinctly moral ends, not distinctly human ends, and if this pursuit results in AMAs who also do accordingly, then those AMAs are anything but anti-Kantian. They are rather morally requisite. Perhaps Superman is an AMA.

[9] Kant details this central role of conscience more in later than earlier writings, though the portrait herein represented is present at least as early as 1785. For example, it is conscience that moves the bad-faith borrower to reason that to abuse the privilege of promise-keeping is to destroy the institution of promise-making, nullifying his will to promise to repay what cannot be repaid under the objective law of morality "since no one would consider that anything was promised to him, but would ridicule all such statements as vain pretenses" (Kant, 1785, page 39). Without conscience, none of this happens.

Versenyi also argues that, not only are Kantian AMAs morally requisite, if we do not pursue their creation, we will be doing ourselves a moral injustice. If we have the capacity to further moral ends in the form of moral robots, then "not to do so would be to neglect one of our natural gifts and this cannot be willed as a universal law of nature" as well as "to neglect humanity in our own person" (Versenyi, 1974, pages 255-256). This follows - in the Aristotelian spirit of course - because natural capacities are to be developed according to natural purposes, including moral capacities to further the ends of morality, virtue e.g. to make things the very best kinds of better by embodying those things necessary to get those very best kinds of things done. But, this is all beside the point. The important thing to note is that it is not at all clear that Kantian AMAs are logico-necessarily anti-Kantian.

Of course, none of the preceding does anything to solve the reliability issue. Versenyi concedes on this point, writing that "the trouble" with Kantian moral agency – e.g. "free" and "rational" - implemented in a "perfectly safe" – e.g. "behaviorally limited" - AMA is that "it will not work as long as we have not also built into the robot an almost infinite knowledge of what is in the long run and in any conceivable situation good or bad, beneficial or harmful to human beings," "not even a theoretical possibility" (Versenyi, 1974, page 257). So, perfectly safe appears to be out of the question. Instead, Versenyi proposes that the "only way" to make AMAs "safe for men" is not by trying to fully prescribe their behaviors, but rather by making them "morally isomorphic", "wholly identical in structure and programming with human beings" (Versenyi, 1974, page 257). The following section will pursue this proposition, where we will uncover what is essential to this moral isomorphism, and so what it means to be “wholly identical” in the morally relevant sense, beginning with a brief review of complementary schemas for moral agency.

Part 2

> Thus, when practical reason cultivates itself, there insensibly arises in it a dialetic which forces it to seek aid in philosophy, just as happens to it in its theoretic use; and in this case, therefore, as well as in the other, it will find rest nowhere but in a thorough critical examination of our reason
> -- Immanuel Kant[10]

2.1

To date, the most influential classification of different degrees of moral agency has been due to James Moor (Moor, 2006, 2007), and is central to Tonkens' assay, as well. Moor's account proceeds thusly. At the lowest level, an “ethical impact agent” is any object the actions of which have ethical consequences. Robotic jockeys in Qatar freeing human jockeys from servitude is Moor's example here. One level higher, “implicit ethical agents” are morally significant by design. Moor's examples here are spam-bots and airplane instruments that warn pilots of unsafe conditions, direct extensions of human moral agency. Moor’s third type of ethical agent, the “explicit ethical agent,” is an indirect extension of human moral agency. Able to identify morally salient information within specific contexts and to act according to appropriate, externally derived, principles, Moor feels that this is the “paradigm case” of robot ethics, “philosophically interesting” and “practically important” while not too sophisticated to be realized. Notably, this represents Tonkens’ target for Machine Ethics, “an ethical robot, not one who sometimes acts ethically, or that can act unethically” (Tonkens, 2009, page 429). Finally,

[10] Kant, 1785, page 22

the “fully ethical agent” is Moor's fourth degree of ethical agency. This is not a level of agency likely to be realized in robots, on Moor's account, representing as it does the characteristics central to our ultimate concerns - e.g. autonomy, rationality, dignity - to date unrealized in artificial agents, and indeed seldom enough demonstrated by human beings.

This raises another aspect of the moral hard problem, the problem of moral ascription. The problem of other consciences. After all, even with Moor’s classes in hand, it may not be possible to reliably classify agents purely on their behavior, and so to reliably ascribe appropriate standing when necessary. Consider the treatment of the indigenous races, from the African to the American to the Palestinian, by the colonial West. Moral status changes by ascription, and for the better when ascribed – or withheld - for the right reasons. This trouble points to an aspect of autonomous agency neglected on Moor's schema, the perception and consequent ascription of autonomy due to demonstrations of agency along the lines of Ziemke’s “phenomenal” autonomy mentioned earlier (c.f. also Schermerhorn and Scheutz, 2004).[11]

How does a moral agent demonstrate morality? Expressed sensitivity to sufferings of similarly embodied others? Ability to independently set and maintain distinctly moral goals? Both? Such an agent would clean the house when so directed - pursuing external goals - and while on that task may stop

[11] Schermerhorn and Scheutz (2004) have set out a three-tiered schema of agency sensitive to this issue of moral ascription. Their first level involves the execution of some function without direct human assistance, as a direct extension of human agency. Robotic jockeys would qualify here, as would robotic vacuum sweepers. The second level involves pursuing more open-ended tasks, without the need for step-by-step direction, as indirect extensions of agency. Military missions would qualify as such tasks, with this degree of autonomy expressed by mission-capable explicitly moral 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, and representing our target at the crux of Tonkens' dilemma, the fully autonomous Kantian AMA.

sweeping the floor to rescue a neighboring family pet from a house fire. Stopping the house cleaning to rescue the neighbor's cat from a house fire invites an ascription of autonomy, and of moral agency, if it is done out of conscientious concern for the well-being of the cat and caring extensions thereto. Not stopping the house cleaning while the cat suffers affords no such opportunity for ascription, unless the robot set the fire. This is troubling for two reasons. For one, without the house fire, without the opportunity to do the right thing at the right time, there may be no recognizable difference. In order to demonstrate moral worth, a cunning robot may cause a fire, and present itself a hero. If attacked it may stage a bombing, and present itself a victim. Rather than the moral zombie, this is the robotic psychopath. And the return of the possibility of the moral monster reminds us of the second reason why burning cats matter. Sensitivity to opportunities to diminish and prevent especially local suffering is the efficient means to the right ends in most cases. Imagine the cat rescuer proceeding as follows, preparing to act, watching closely from a near location, communicating with witnesses, investigating the conditions of entry while contacting fire departments, noting that the fire's location is remote from the cat's, reaching a hand through a window, easily attracting and then removing the cat to hold it safely until a suitable caretaker presents itself. Demonstrations of autonomy are dependent on opportunities for exercise, as well as trained capacity to exploit them. If the robot had not dedicated some time and energy to establishing good relations with that cat during months and years previous, then the cat may not have come so easily, and the rescue may have not been so easily effected. Finally, the opportunity to do the right thing only arose from an effort to establish that potential, perhaps by – at least for a moment, every morning – operating outside of direct commands from owner/operators to let the cat lick the egg from the breakfast plates before washing them.

A big part of autonomy is also *not* pursuing certain opportunities. Saying “no” to the house fire idea to rid the world of a worrisome cat, though it lies in the agent’s potential, is an example. Another part is the opposite of any apparent, autonomous initiative whatsoever. It is doing as one is told. Going along with. Doing as one is told would seem to be demonstrated evidence against autonomy, not for it. However, doing as one is told is often the right thing to do. Fully autonomous agents often do what they are told to do, pursuing objectives solely of external derivation, freely ceding autonomy through “transfer-of-control” (Pynadath et.al., 2002), e.g. as part of a team, family or even kingdom. This capacity to freely “adjust” individual autonomy to suit a given context is essential to autonomous agency, not evidence of its absence. Without this capacity, to give one’s self over to others, as well as its converse, to say “no,” there can be neither autonomy nor morality. All of this confounds any easy ascription of autonomy on phenomenal bases, as well as raising questions about the presumed moral status of existing moral agents, e.g. human beings.

For instance, in the great team that is the military, both robots and humans are to be equally 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). As a link in this chain of command, any failure to follow orders is not revered as moral virtue. Rather, moral autonomy – especially the power to say “no” - is condemned as malfunction or worse. The soldier gives up his self-sovereignty, and with it ethical liability, remaining autonomous only the mode of optimization for external goal achievement.[12] In so far as we aim, as ethicists, educators and engineers for this as an ideal, then we are aiming for something else besides moral agency, something "separated from law and justice." Imagine merely that a Kantian moral agent would certainly suffer moral self-repugnance at murdering women and children at a wedding party, then "double-tapping" their rescuers – clearly military robots are not Kantian aims.[13] No, the target agent is simply a kind of efficient cause for its correspondent situation, and Machine Ethics taken merely as more-or-less malleable means to its realization. In the face of the momentum of technological progress, Beavers' fears of "ethical nihilism" may have already come true. If this is the end (goal) of ethics, then this is the end (finish) of ethics. Nothing to discuss here. Time to move along.

Interestingly, Kant also warned against the "quiet death" of morality through the reduction of autonomy to "the physical order of nature" (Kant, 1780, page 7). This is to say that something vital may be left out of its articulation, leaving us to mistake *what* (apparently) *is* for *what might be* or worse, *might have been*. The clue to this absence is in the moment of

[12] This is debateable, and I have made the case elsewhere that the soldier never fully cedes conscience, with the proof in the consequences. The tragedy of this exchange is that the soldier typically suffers after the missions are complete, and they reflect on their actions. In the USA, many veterans are killing themselves daily. There have occurred recently a number of shootings at military bases. One young man killed his officer and himself, because he was denied an opportunity to remove himself from offensive conditions. He needed a break. Didn't get it. Broke, instead. Others feel guilt at murders, and worse. Others, still, are so betrayed by their government, that they have no homes, no caring support structure to aid their return to civil society after, in many cases, five or ten years of forced service in morally abhorrent situations.

[13] Consider similar stories here http://www.bbc.com/news/world-us-canada-24557333, here https://www.commondreams.org/headline/2013/08/01-4, here http://www.theguardian.com/commentisfree/2012/aug/20/us-drones-strikes-target-rescuers-pakistan, and even here http://edition.cnn.com/2012/09/25/world/asia/pakistan-us-drone-strikes/index.html.

discovery that one could have done better, in feeling the regret at inhabiting a situation that is deficient due a deficiency in moral sense and judgment. What seems to be missing is autonomous agency in the self-sovereign sense, represented in the neglected promise of continued research in artificial intelligence – a fully autonomous AMA. And most importantly, it represents the promise - indeed the aim - of traditional *human* moral education, autonomous human agents (AHAs) thriving in a just world of their own creation. This is the sort of agency that we are *supposed* to create, regardless of mode of embodiment, agents “morally isomorphic” in this basic way, self-sovereign and invested in the future as the sole domain of personal residence, aiming to make that world better in the administration of justice.

Raising the perfect political animal is an especially weighty obligation for new parents, and exactly the one that some commentators would have us shirk. But, should we pursue it, if virtue can be taught – designed, engineered, entrained – then its reliable reproduction will benefit from some clarification of constitutive structures and processes, necessary ends and means. Indeed, making morality explicit in this way has been, for me, the single most compelling reason to research artificial moral agency, consonant with Versenyi’s pro-Kantian argumentation. The aim now is to expose the dynamic structure that underwrites Kantian moral agency, as it will serve as the comprehensive account of morality common to human and other moral agents.

2.2

In this section, we will understand the essential structure of moral agency, and how this may be realized in natural as well as artificial agents, beginning with the embodied logic of natural purposes and that material capacity denied AMAs on Ziemke's (2008) account, autopoiesis. We will then search for the phenomenology of this process, and find it in Kant's conscience, returning us to topics of autonomy, self-sovereignty, and dignity with which this paper began.

"Autopoiesis" comes from the Greek meaning "self-producing," with "allopoiesis" its contrary. The "primary function" of an autopoietic system is "self-renewal through self-referential activity" with an "allopoietic system like a robot deriving function from an external source" (Amoroso, 2004, page 144). "Autopoiesis is also the mechanism that imparts autonomy to the living" (Luisi, 2003, page 52), with "the minimal form of autonomy" in these terms being "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). This minimal self-renewing self-referential system set apart from the external environment is self-organizing, "one that continuously produces the components that specify it" (Varela, 1992, page 5) through processes that require that it maintain itself far-from-equilibrium with that external environment, e.g. little animal Carnot engines. Though ongoing, these processes in their envelope constitute a "concrete unity in space and time," with the target character of this unity itself making "the network of production of components possible" (Varela, 1992, page 5), i.e. it becomes some target something, self-optimizing.

The work required to achieve some ideal condition is facilitated through metabolic storage and "budgeting" of matter and energy, and these resources constitute the "bodily fabric" of the unity (Boden, 1999, 2000). The bodily fabric emerges as a single, bound entity within "molecular space," with its properties [14] "structurally determined" by potential and actual chemical changes to the system of internal and external environments moderated by the selective permeability of the "membrane" through which the dynamic balance between the two is selectively moderated (Romesin, 2002). Together, agent and environment constitute a "niche" co-emergent with the exercise of agency itself, i.e. action changes the world, and the changing world changes the agent, and the niche is that fit with the world on which the processes of the entity in question depend. This niche represents the "operational closure" of the autopoietic 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). As the autopoietic system acts, expresses, and creates itself, it ultimately creates its own world (Luisi, page 58). It then depends on this world of its own creation for the maintenance of its continued unity. It is a "self-producing coherence" that must actively maintain a stable niche, or fail to remain a unity far-from-equilibrium, becoming instead dissolute and dissembled, i.e. dead. In systems terminology, the autonomous agent is 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).

[14] Including semiological properties, as especially on this view signs and symbols are not abstract tokens but rather material tools.

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 procedural unity of the system is maintained, and “disintegrative changes” through which it is not. “Operational closure” extends from the molecular to the modular (e.g. cells and organs) to the systemic (e.g. the organism, whole) levels of organization, and upwards to social, cultural, and philosophical levels including a prospective agent’s aspirations and ideals in organizing all of these levels in the most beneficial ways for all concerned in maintaining similar unities and their dependent processes. From seed to salvation.

In this way, then, an agent’s niche can be understood as layers of increasing conceptual order established in pro-active defense against disintegrative change. Autonomy is the capacity to more or less determine the character of these layers of defense. Where these defenses fail there is the loss of integrity, and the capacity to selectively integrate with the external environment.[15] In so far as established orders resist spontaneous devolution, they can be said to last forever. And, to create such orders through autonomous direction of resources specific to the bodily fabric is the aim of inquiry, as Aristotle told us - *nous*. Where these resources are directed toward more or less ideal organizations of agents and environments, they affect the niche in terms of which subsequent generations will be informed. These changing environments are the levers of evolution, creating – as in the contemporary case – artificial bottlenecks in survival opportunities, but this is beside the point. The point is that the bodily fabric constitutes the current as

[15] Death is a busted membrane, and these membranes can be burst in many ways, e.g. with pointy sticks, with betrayal, and even with broken promises.

well as the future dimensions of necessary interest within any given situation, essentially shaping any possible niche. This cannot be escaped – this is what it feels like to be embodied. “Our minds are, literally, inseparable” not only from our bodies but from the environment as we experience it – temporally as well as spatially - thereby constituting a peculiar sort of “prison” (c.f. Rudrauf et.al., 2003, page 40). And perhaps it was for good reason that Kant advised earlier that we visit the poor and imprisoned, as it is here, imprisoned, that we have uncovered the comprehensive grounds of Kantian moral autonomy.

The essential codetermination of agent and niche definitive of operational closure represents a process easily correlated with the Kantian portrait of self-sovereign moral agency as described in the second section, and most specifically as laid out in Kant's "kingdom of ends" characterization of the categorical imperative.[16] Only the free agent can be imprisoned. Being in a prison is being necessarily attached to a situation, *en*-niched. A free agent can determine, more or less, the situation to which it is attached.[17] The operational closure of the Kantian moral agent extends to the kingdom of ends over and within which the agent recognizes itself as both law-making authority and bound subject. Reciprocally codetermined with its effected niche, thus, the measure of agent autonomy (if not the measure of agency,

[16] "A rational being must always regard himself as giving laws either as member or as sovereign in a kingdom of ends which is rendered possible by the freedom of will" (Kant, 1785, page 50).

[17] This is why “religion” means “re-“ “-ligion” or “bound back to” and “freedom of religion” means “unhindered capacity to determine, for one’s self those ends which tie back to one’s self, and that pull one’s self forward toward that ideal therein represented.” Heidegger refined Kant’s theory of Aristotlian political agency in the simple resolution that it is of the essence of *dasein* to “be there,” or more correctly “be the there” of which the kingdom of ends is only the ideal against which one realizes himself as a “not” – as less than. That there is one situation, notably a situation without a situation, that characterizes dasein and that does not yield to graduation. That is death. And, it is the face of death, not-being the situation, that human beings turn to construct and to value orders that outlast them. It is also in the face of this certainty that others value destruction and disorder. Death is the one situation that is not within the mortal man’s capacity to affect, or effect, either way. And this is why religion is associated with Gods, representing those capacities to establish and maintain orders that are naturally prior to and beyond the human, while to not work toward these orders is godlessness.

itself) is ultimately the degree to which the world that an agent creates through its actions supports the exercise of that same creative agency. As all creative agency is essentially autonomous agency, and all human beings enjoy the potential for autonomy, the target niche includes “mankind in general” and is equally expressed by “intelligent being” or “rational being” in general. The kingdom of ends, thus, is not simply one possible end of autonomous agency, it is the necessary end.[18]

Let’s look more closely at how and why this end is, and is to be, effected. Consider two further concepts from the autopoietical lexicon, “homeostasis” and “decoupling.” “Homeostasis” (or better “homeodynamics”) names the balancing of internal and external forces through largely automated physical processes, and has been fruitfully developed most famously in the cognitive sciences by Antonio Damasio. On Damasio’s account, the play of these processes constitutes the "feeling of what happens,” e.g. consciousness and mood, and extends from highest to lowest levels of organization and back again. From the bottom, the cells of the body “gravitate” toward “fluid” states and away from “strained” “configurations of body state,” and these material dispositions 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). Processes efficiently rendering unstrained configurations are generally enacted, and where these processes

[18] In sum, as the "rational" moral agent acts in pro-active defense against disintegrative change - in order to remain rational, self-sovereign, autonomous, not a slave in perpetual “nonage” – he/she/it is structurally bound to the kingdom of ends, free to act toward this end, and this end only, otherwise not working to perpetuate the very best aspects of that agency, its rationality, its autonomy, its morality. To violate the categorical imperative is to act contrary to the natural purposes inherent in the embodied fabric. It is a niche-breaker, the Kantian moral nutshell. Cracked.

are hindered or helped, implicated objects and relations are devalued or even over-valued in corresponding ways. Thus, the positive association with objects that facilitate stability, and the negative with those that threaten it, transform 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” corresponding as they do to “some structure of the body, in a particular state and set of circumstances” (Damasio, 2003, pages 197 and 56, respectively).

This description certainly captures the sense of a "composite feeling" of “integrated being,” and presents a comprehensive basis for moral agency with relevant processes common to both artificial and to “living” systems.[19] The feeling is of “gravitation” away from strained projections and toward unstrained projections, away from disintegrative changes and towards integrity. Transitioning between relatively strained and relatively unstrained embodied states, the agent envalues situations and salient object engagements therein. “Good” and “bad” are tags like tastes of situations, due at least in part perhaps to the hippocampus and its central role in the especial spatiality of memory and of memory formation, especially as informed by the sense of smell, which is a distinctly local, intimate, carnal sort of knowledge about the world. The fact that the sense of smell is so important to memories and their formation, the shape of memory is essentially chemico-physical, as the information itself is the thing itself, is its effect on the chemical consitutents of the organs of sense, themselves, as they are excited, and driven away from sleepy equilibrium, with their tremors sent up the lines to their witness, and

[19] Though, admittedly in the former these remain simple.

the change of situation is registered, taken as a whole – given a flavor, good or bad – and compared with others, possible, or lost.

One may argue that this is due to the biological embodiment of the human being, a side-effect of the hippocampus from its early operations. But, the reverse is true, in my opinion. In my mind, the essential spatiality of human memory is due to the nature of agency, due to the spatiality of the experienced world as evaluated by a point-source agency with limited potential in limited dimensions to act over that encompassing terrain towards the best possible situations. This spatiality extends from physical to metaphysical, and is why we are able to say – for example – "You need to keep yourself up" or "My car broke down." It is the up of order, and the down of decay, that characterizes not only the natural cycles in which our evolution has left us enwombed, but the resulting embodiment, as well. We put the good stuff in the front, up top, bad stuff comes out the bottom, and back behind us. Neurons at the very top of the head continue to develop, as we grow old we continue to grow up. To be is to be good. But that is all beside the point. The point here is that any number of artificial systems may be envisioned to operate according to this embodied logic. We can project a portrait of agency that might be consistently applied in the evaluation of either sort of system, artificial or natural. Indeed, it can be applied in the description of any material system, at all, and this opens a window not only on Kantian moral agency, but on the notion of objective moral law.

For example, consider homeostasis in terms of the integrity maintaining material system par excellence, the molecule. The nominal representation of any given molecule, such as that in a chemistry handbook, is that of a system of self-interested centers optimally arrayed with lowest strain, i.e. static and at lowest internal energy . However, this nominal representation is itself only an ideal, as it presumes that the material exists in an ideal situation, at zero degrees Kelvin without external interference. *In vacuo* with only the electronic energy of communication within and between its constitutive “organs” remaining – the energetic sunk-cost of being a thing - this is the closest that a material system like a molecule can get to autonomy. It determines its situation because it is the situation – there is nothing else. Imagine such a luxury afforded something like a rock: finally free, self-determining, a space all its own, self-sufficient in the fullest sense, such isolation from the material and energetic background is an ideal condition, only logically possible, i.e. we can conceive of it. The trouble is that this picture is fiction. The everyday molecule is a busy servant to external forces, forever a slave. A more realistic image of a molecular system oscillates from strained configuration to strained configuration in dynamic equilibrium between forces internal and external, passing only briefly though ideal arrangements, if at all. What we experience is the system in situ. The idealized version appears in textbooks because it is simple, both easily reproduced and, because of the nature of its mutation, uniquely informative. In its deviation from reality, it does represent that condition towards which all natural systems can be thought to aim, an ontic ideal and – bear with me - Plato’s form of the “chair.”

Now, a molecule might not “create its own world” in the same ways that a moral agent does, but its presence does influence its environment, and this influence can be bootstrapped into a good explanation for the evolution of higher-levels of organization, “life” and all of its constructs. In special cases, the "gravitation" toward stable states that is essential to all things in nature grounds the emergence of the cellular and then organismic levels of organization from the atomic to molecular and up. As complexity increases, it is the energetics inherent in these self-organizing processes that provide for the seamless transformation of the merely physical world of objects into the metaphysical space of value. This is why we may say of a chlorine ion that it is "hungry" for electrons, as it makes sense of the space within which the ion must balance internal and external forces in procedural terms similarly captured in our own embodied systems when empty, deficient, and so motivated toward wholeness and felt unity. The difference here being, simply, that the chlorine may remain satiated forever, while that is not the mortal human realization of this condition. Our needs are many and diverse and multiplying like a disease.

Regardless, the molecular logic common to all natural systems ultimately gives rise to the "the feeling of what happens," with even social systems emerging from roots in "molecular space." 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. Accordingly, we may suggest that all systems - including "artificial" systems - seek equilibrium in terms of their environments, and understand homeostasis as the tendency for some dynamic systems to compensate for forces of change while maintaining stability far from equilibrium from the immediate environment, *en*-niched.[20]

[20] What is left for ethics, thus, is to identify those terms defining of the situation that do not contribute

As we have seen, niches are spaces of cognition and action, and these are optimally protective against forces of disintegrative change. The capacity for certain complex organizations to set the terms of the co-constitutive realization of self and world towards which they are motivated is the autopoietical foundation of autonomy, the prison of freedom.[21] This prison of self-protective self-optimization is fundamentally realized at a specific level of organization, in Luisi's "living" molecules, micelles. A micelle is a niche par exemplar. Its structure insulates its interiority from upsetting external pressures, constituting a fundamental integrity partially "decoupled" from the environment. "Decoupling" means "reducing direct effects of environmental stimulation and opening up possibilities for internally regulated behavior" (de Bruin and Kastner, 2011, page 10). So decoupled, a micelle is a proto-semiological bubble of self-production, effectively creating its own world within itself and of its own resources. This capacity to decouple from the environment is an essential aspect of autonomy, freeing an agent to act according to internal constraints rather than reflexively according to external triggers. This freedom is especially refined in human-level, Kantian-rational agents, as internal constraints extend throughout the range of agency, from chemical to symbolic, from here and now to the kingdom of ends, with capable agents creating their own purely conceptual worlds from their own cognitive resources, e.g. symbols, formulas, principles, maxims. Moral micelles.

to maximal efficiency of this balance between internal requirements and external requirements optimizing for salient qualities of life. I am sure that killer robots will disappear if the future were viewed through such a lens.

[21] This is essentially what I mean by "self-abduction."

Moreover, decoupling from the immediate environment allows a rational agent to weigh one total system conformation and corresponding condition, one niche against another, enabling autonomous action toward either, facilitating “hypothetical thought.” And hypothetical thought isn’t limited to just one situation for comparison, rather any situation may seem possible. One way to facilitate rapid comparisons of possibly embodied situations is through the recursive combination of formal constructs including counterfactuals and imperatives (Stanovish and Toplak, 2012). This capacity for formalized hypothetical thought allows an agent to reflect on Kant's categorical imperative, testing actions for contradictions of will, identifying characteristic marks of target situations. This capacity to formally represent alternatives that guide thought and action, for example as principles or as commands, can be understood as a special sort of “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). In the human realm, this formally mediated “incoherence” grounds the emergence of social and moral systems non-descriptively represented in theories of ethics and writs of history and law, i.e. "oughts" and "shoulds” not “isses” and “weres.” 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 the future” (Beer, 2004, page 324). This capacity to decouple from external pressures through symbolic mediation and to coordinate action to commonly beneficial ends which may only arise beyond the individual agent's felt spatial and temporal limits is a powerful evolutionary force, known in traditional moral theory as “freewill” (c.f. Juarrero, 2009).

Recalling Aristotle’s characterization of the perfected political animal as discussed in the first part of this paper, and recalling Tonkens’ original challenge regarding an agent both rational and free, it is important to note, here, both that the combined capacity for internal reflection over, action towards and public representation of projected ideals constitutes political agency, and that, in the proper, unimpeded function of these capacities, ideal situations – the globally optimal niche – are sought after and requisite processes enacted, relevant virtues embodied, and kingdom of ends made real. The mode of embodiment is beside the point. Of course, AMAs can be moral. Properly configured, they must be. How reliably will depend on how thoroughly we corrupt them, as we shall discover in the next section.

2.3 Autonomous reboot

We may seem to have wandered off from our point of origin. After all, Tonkens mentions neither autopoiesis nor syntax autonomy in his assay of Kantian ethics. Yet, these concepts can be consistently incorporated into the Kantian manifold. Indeed, Kant anticipated the autopoietical distinction between life and artifact in terms of self-organization. For Kant, in both a living thing and an artifact, each part exists “by means of the other parts” as well as “for the sake of the others and the whole,” while only for the living thing are its parts “all organs reciprocally producing each other,” so constituting “a whole by their own causality” (Kant, 1790, page 202). Accordingly, a living 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). Life has purpose as reflected in its organization, an end in itself. Rocks, stones, sticks, slaves – these do not, but only in the case of slaves is it wrong. This is because a living thing 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). Thus, life seems to contribute an essential moment to Kantian moral agency. It represents the capacity not simply to do, but to become the one who does.

Life also represents something else, something more, something separated from justice. Kant puts the matter thusly: "Life is the faculty a being has of acting according to laws of the faculty of desire" (Kant, 1788, footnote page 9). That is, it pulls towards the "kingdom of trends" rather than the kingdom of ends. Subjective, arbitrary, this faculty - desire, bodily inclination, the press to satisfy immediate and local material needs - operates as a necessary counter-pole to moral duty, the conscientious, rational entrainment of self to the objective interests of "mankind in general," the kingdom of ends rather than trends.

Still, both are necessary. Without tension between the two, between immediate feeling and rational absolute, there can be no Kantian moral agency. It is in the movement from one to the other that moral agency demonstrates itself (given the opportunity), as it is in this movement that morality itself becomes real. Moral agents aren't created, any more than they are born. Moral agents become moral in "the administration of justice," the first step being the fulfillment of "the one duty that there is room for," i.e. moral education. So informed, rationality counters immediate material inclination, desire, and "guides action in accordance with objective laws, towards the end of establishing a good will and moral character" (Tonken, 2009, page 426). That said, though material inclinations may differ, nothing here contravenes an artificial agent embodying the same motivational logic, and even being admired for it.

Some may contend that “living” means not-artificial and if we create living robots then we have instead not made intelligent machines, but dumb flesh, and that in any case a Kantian agent is partly dumb flesh. But, I think that there is room for the artificial non-living (in some organic sense) Kantian moral agent, and will spend the most of the rest of this paper developing this potential. Consider, for instance, that Kantian "rationality" is purposefully broad, and that autonomy, “autonomy of the will,” or “freedom” as he variously calls it, “is a property of all rational beings”: to be free an agent must merely “regard itself as the author of its principles independent of foreign influences”(Kant, 1785, pages 64-5). Self-sovereignly. Of course, “foreign influences” include “the faculty of desire,” but are certainly not limited to human-only equivalents of it. The point here is merely that rationality requires that there be morally constitutive tensions between material projects and moral duty. These tensions can and do arise by way of human embodiment, but are certainly not limited to it. All that is necessary is that an agent must act from moral duty, that he/she/it must do so from his/her/its own resources, autonomously, through the self-determining identification with all those similarly motivated. Any further condition – like a natural birth - is beginning to look like prejudice, or perhaps much worse – anti-Kantian.

In light of these results, let's reconsider Ziemke's distinction between "noumenal" and "phenomenal" agency. 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). We can understand "immortal" as free from all of the motivating necessities of embodiment, including the drives to maintain bodily integrity that so occupy the living agent, e.g. decoupled, syntactically autonomous. We can imagine immortality inhering in mathematical truths. We can also imagine an immortal condition, for ourselves, arising in one of two ways. First, the environment is such as to meet all needs easily, and second the body has no needs to be met. If the body has no needs to meet, then all ends are met. Dead. On the other hand, if the environment meets all needs, then this is merely the view from within some ideal end, agent and environment, coupled. As only the needful can have needs met, it follows that for the needful this ideal end represents a categorical pull forward, to meet their needs. This pull is internally active, not dependent on the capacities unique to any supreme or individual entity. It is universally motivational, potentially immortal, again Plato's form of the chair, but here of course best recognized in Aristotle's unmoved mover. This is Kantian moral religion, and may be the basis for a robotic one.

Gravitating to ideal ends, with tension at turning away, this is the "what it feels like" to be a Kantian moral agent. Not motivated "to" some end, but motivated "from" some end. Not according to duty. From duty. The ideal pulls the moral agent forward, only and always forward, so that whatever is done under the sway of that ideal is the right thing to do. This is how we know that anything so motivated might be engineered to always only attempt right actions at the right times. There is no crime without malice, negligence, or dispassion entirely. Nothing here contravenes an artificial Kantian moral agent, even a perfectly reliable one, though obstacles in resolving artificial goodwill remain.

For one, the Kantian autonomous agent doesn't simply act independently of all material dependency. Kantian autonomy is not the radical freedom we may mistakenly imagine syntactic autonomy to represent, i.e. constrained only by logical possibility. It is also not the freedom to act in a purely detached, scientific way, to "muddle through" difficulties as they present themselves. Such associations with the ends of action are merely contingent, while for the self-sovereign Kantian agent there exists a relationship of necessary and lasting residence, such as that between king and country. This sense of codependence provides for the difference between the noumenal and the "archetypal" worlds in the Kantian theory (Kant, 1785, page 44). The difference is felt, but it is also functional, and so potentially detectable. The perfect epistemic agent may seek the noumenal, but the perfected political animal, the Kantian moral agent, seeks the archetypal. In pulling these two apart, we may come to a phenomenal measure of moral agency, with some surprising results.

The archetypal world, previously given as the "kingdom of ends," is also referred to as the "summum bonum," the "supreme independent good," and even "God." As alluded, prior, "archetypal" differs from "noumenal" in that it is the one situation with which the 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,” but bears deeper resemblance to Socrates' projected plane of justice to which he freely binds himself, and in personal relation to which is guaranteed to not do wrong (c.f. White, 2012). Kant explicitly rejects the notion that recognition of any particular “God” is necessary for autonomy – and with this goes any explicit requirement of a Cartesian “soul,” as well (Kant, 1788, page 133). Rather, the important relationship between the moral agent and the kingdom of ends is that the agent “own” it. He/she/it must intend to reside in its terms, as that situation arising from his/her/its own agency.[22] Sleep in the bed you have made. And, autonomy, moral agency, and political agency are essentially capacities to determine the condition of the bed.

An agent need simply understand this fact of its existence in order to understand its moral freedom to pursue the kingdom of ends through the individual life of political action. Accordingly, Kant argues that autonomy merely requires three self-conceptions active at the demonstration of agency: freedom (specifically, conceiving one’s self as having the capacity to self-legislate), immortality (conceiving one’s self as if unbound by physical and temporal constraints on the preceding), and God as the existence of a “supreme independent good” (Kant, 1788, page 137). This good is the aim, the destination, the archetypal end of action and “object of a will morally determined” in accord with the ideal moral situation. Accord with God results in a deep moral pleasure subjectively realized as “harmony” within the agent,

[22] Synthetic a priori. So, every move counts, and it counts not because of the particular move being made and its particular ends. It counts because the structure of moral agency requires it.

a harmony that we may imagine to be constituted, procedurally, by the constructive overlap across the extant realm consisting of all intelligent beings sharing in this ideal. Resonating with the universal archetypical optimal. There is nothing here denying a Kantian AMA, unless we have arrived at "the end of ethics" as a matter of choice. God may be dead, but the memory remains useful, and robot religion not far off.

2.4

Though Kant's moral theory may fail to account for less-than fully autonomous artificial moral agents engineered especially for less-than full autonomy, it should come as no surprise to find necessary resources to solve the problems of fully autonomous AMAs. After all, Kant himself worked toward a similar resolution, asking "What, then, is really pure morality, by which as a touchstone we must test the moral significance of every action?" (Kant, 1788, page 157). This paper has so far established a preliminary answer to this question inclusive of artificial agents in their conception, deriving from the two-sided nature of Kantian moral agency, namely the motivating harmony established between the moral agent and its target environment, the "kingdom of ends." This final section will set out the final piece required to answer Wallach's call for a comprehensive "platform for testing" moral agency, and will finally dispel Tonkens' apparently insoluble dilemma blocking the properly configured AMA's claim to space within the extant moral community, before considering implications for the aims of Machine Ethics going forward.

Tonkens, it may be noted, limits his analysis of the dual aspect of the Kantian moral agent to its necessity in the constitution of moral virtue through the exercise of reason in the conformation to moral duty, writing that "According to Kant, it is only because humans can violate the moral law and succumb to the temptations of sensual satisfaction that they can truly be said to be moral agents" (Tonkens, 2009, page 426) and noting that "According to Kant, part of being a moral agent means possessing ‘‘the capacity to master one’s inclinations when they rebel against the [moral] law,’’ hence the ability to freely commit actions that are not moral" (Tonkens, 2009, page 430, quoting from Kant's *Metaphysics of Morals*). It is also worth noting that Tonkens employs this duality to stipulate, once again, that the "goal of Machine Ethics … is precisely to reduce morality in robots to something like unchallengeable inclination" (Tonkens, 2009, page 430). Of course, were this truly the settled case, then there would be no need to pursue this discussion, no need to issue any "challenge,” and indeed, no need for Machine Ethics, at all, as it may as well be called the “Ethics of Hammers.” But, how can we understand Kantian ethics in a way that finally obviates these concerns about agent-level moral reliability?

The key here is in the nature of the motivation from moral agent to moral ends. As we have seen, there are two sides to the Kantian moral agent. One side is bodily, embedded, source of heteronomy. The other is decoupled, rational, source of autonomy. One side is product of the immediate environment and of "animal" attraction to “objects of volition” within the phenomenal world of sense, while the other is “immanent” through “transcendence,” committed to a “world of intelligence” and the product of “intelligible being” (Kant, 1788, page 108). These two sides constitute two essential poles within the Kantian agent, one material and one ideal, and from these positions each 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). These two points of view inform that capacity for the self-sovereign autonomous agent to “pass judgment upon himself and his own actions,” the structure that as a functioning whole is the “conscience.” Let’s look closely at how this process plays out.

In projecting the highest possibility, the agent decouples from the immediate environment, “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,” to be a “better person” and to be "a member of the world of the understanding” (Kant, 1785, page 72), i.e. as revered sovereign over the kingdom of ends. From this exercise, he is able to self-legislate – he must conceive of himself as free in the self-sovereign sense. Maxims arise autonomously – “free from all laws of nature” (Kant, 1788, page 118) - as principles directing action toward this ideal inner harmony between one’s self as material peasant and kingly project.[23] At this point, autonomy is possible, as this structure of agency provides for actionable alternatives. When there exists harmony between agent and ends, along dimensions of action bridging the two, action may be initiated and ends may be achieved. In this exercise, autonomous moral agents become themselves in situations more-or-less of our own making. These selves are more-or-less consonant with some ideal

[23] Yet, autonomy is constrained, as principles and possible ends are "determined by predicates of [his] own nature" and enacted consistently with the laws of nature (Kant, 1788, pages 141 and 118), i.e. consistent with capacities for agency.

projection, some anticipation. The result is weighed against the projected optimal. And it is this information, the self-identification between an agent and its ideal end (its global motivational intension) indicated in what counts as error for its correction that might provide a reliable indicator of moral system performance regardless of accidents and acts of God. This tension is the measure of good will. Counting the right kinds of errors a sign of a functioning conscience.

In so far as an agent is conscientious, the kingdom of ends really is real, here and now, as it is the always current, constant ideal against which all situations, beginnings and ends are compared, and according to which actions are measured for their progress or regress thereto and therefrom. A robot zealot (if this is not redundant) can be imagined to operate under an especially steep slope of inclination toward some ideal state of affairs, feeling deeply urgent about the distance between the idealized optimal and the suboptimal arrangement of all things within its enacted environment. All things in this metaphysical headspace take up a weight due the "gravity" of this polarization across their containing fields. Every object met with is envalued for its capacity to facilitate global ends, and personal capacities to employ tools toward certain ends are developed. This is the identification of and entrainment towards global moral goals. The cultivation of resources to solve moral problems. The cultivation of conscience.

Note that this process depends directly on the capacity for a rational agent to place itself at the end of action, and to weigh projected ends in comparison "as-ifs." Autonomy exists in the capacity to determine what these ends are, directing action from the end, first. "Moral pleasure" is merely the harmony established within a rational being as it places itself at moral ends, arising also in the consideration of such ends as he/she/it “transfers himself in thought.” Note that this pleasure extends from the origins of action. It does not depend on consequences. It is of the prospective “mind” of the agent to realize only the ends of “mankind in general.” Harmony comes first. Goodwill, incorporated. Perfectly reliable operation.

With this we have secured the moral status of Kantian AMAs. These are agents which are essentially moral, and so deserving of moral “rights.” Kant writes that, 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” (Kant, 1788, page 70). In exemplifying that process in the manifest of goodwill, described above – freely according to internal resources – the agent becomes moral, indeed manifesting morality in doing so. It is no longer an AMA, it is morality, itself, in demonstration. It adds to the world, freely. So dignified, the AMA is an artificial moral angel.

“Dignity” stems from the Greek, *dekhesthai*, “to accept,” and more commonly means “honorable” and “worthy.” [24] We can understand it autopoietically as arising from the self-referential capacity of a unified and unifying system to accept (and to reject) itself as inseparable from the niche arising from its own actions and processes. To become, or not to become, *that* end of action. Through this distinctly moral cognition, decoupled and unfettered, "free" yet structurally bound as first member of an archetypal world, an agent becomes moral, and it does so along with the moral world of

[24] http://etymonline.com/index.php?term=dignity. Accessed 28 March 2014.

its own enaction. We will know the moral robot by its works, because we should live amongst them. And, should these works express embodied self-determination in virtuous respect for the whole, of which one is merely a constitutive part, then these morally self-constitutive actions enabling the moral world deserve the dignity afforded such by similarly minded individuals, dignity afforded in recognition of inner accord between constitutive poles of moral agency. Inner harmony with ideal ends made obvious is dignified. Resolute, by Heideggerian terminology. As if immortal in the face of death.

This is something more than the blanket equivalence of right action. Dignity deserves reverence. An agent is worthy of reverence when it is deserving of the respect and admiration cum emulation of a beneficent king, and self-reverence in so far as it discovers itself in harmony with the kingdom of ends, i.e. a good king in his/her/its own eyes. This self-regard is the keystone of Kantian moral theory, and its inverse provides the universal gravitation away from immorality. This is the potential for moral self-abhorrence, discussed earlier, the moral self-abhorrence that Tonkens' would save the AMA from suffering. Here, we find that this capacity for moral self-abhorrence is not only unavoidable, but it is necessary to motivate the moral education. Kant writes “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) From goodwill, to dignity, no wrong can be done.

Finally, we have uncovered the motivating bedrock of autonomous agency, and the motivation behind Kantian ethics as well. The “fluid state” of a system attuned to the best conceivable situation for self and others similarly structured motivates (“gravitates”) the autonomous agent to realize that situation from its own resources. The very possibility of morality arises in this rational self-identification, and Kant ties the survival of morality to the realization of its corresponding “pleasure,” that being the low-energy state – “harmony” – that results. Note that he does not tie the survival of morality to the survival of human beings. He ties it to a mode of being that human beings, to this point, have been almost singularly afforded. 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 judgment over ends of action is in every case conscience, and conscience is merely the name for what we now understand to be processes inherent to “ends in themselves,” self-referential, self-organizing “natural purposes” from which emanate the gravity lines of moral principles and precepts that guide actions towards ends within which every moral agent is able to realize a similar harmony, i.e. justice. Any agent so motivated earns its stake in those ends – and deserving of dignity similarly afforded - through actions within and for the sake of this moral community at large, artificial, natural, living or otherwise. Conversely, any agent motivated to the contrary has no such place, and any claim to such community is lost according to its deficiency, until inverted it becomes the worst of animals.

We are now in a better position to guarantee dignity through moral autonomy in a properly configured AMA. For the AMA, itself, there is the index of its unified association with the morally ideal, archetypal world, represented as energetically relative states of harmony and discord with a motivational "pull" toward harmony in relevant dimensions. Note that whatever this ultimately "feels" like is not a question. It is merely the capacity to embody this condition – not the actual feeling of it at any given moment - that gives autonomy to the autonomous. This condition is as a spring, sprung to variously moral ends, and the tensions of springs can be measured. These tensions can be represented in an artificial conscience, as evaluations over past and possible actions and ends. Where ends present situations which resonate with established ideal primitives, then those ends open to further resolution. Otherwise, they are forbidden. This is the motivational moral conscience of Kantian agency. Moreover, conscience may present an explicit indication of the consonance within the agency between projected and ideal ends as a light, turning from blue to red as agents consider over actions from ideal to opposite, and a numerical indication of moral performance may be displayed over the breast as measure of actions undertaken according to internal representations and feedback. We may, in our limited ways, afford the luxury of dignity when all indications show that these agents are working to make our world better in their own limited ways. The robot may be permitted to cover his instrument panel, act autonomously, though with information about agential intentions that is far more reliable than that available from human beings. This robot's morality is not only guaranteed, it is self-evident. One can directly check it.

Agent reliability under constraint of afforded dignity is further guaranteed by the harmonics between agent and ends due to the fact that, once a moral agent harmonizes with the kingdom of ends, this is a relationship that he/she/it is loathe to let go representing as it does the agent-level global attractor according to which all others are evaluable, and away from which any system under the sway of its gravity cannot easily move. There is no incremental release of categorical imperatives – it is a lie, or it is not. Indeed, the measure of the compulsion necessary to overcome the pull of the harmony with the kingdom of ends may be understood as the direct measure of dignity belonging to that agent. Tragically, this quantity had heretofore been hidden, only revealed under the pain of torture required to break a man from his God, and then amorphously, unreliably.[25]

Talk of torture, recalls fears of the ethical nihilism with which this paper began. Beavers' "moral nihilism" is a symptom, a shade, an expression of the deeper incapacity of a productive agent to self-identify with the kingdom of ends. It is the reduction of the traditional moral world into mechanism over and within which none are subject and sovereign, only economic primitives prevail perhaps. For any autonomous being, and indeed, all moral beings, this is the worst end of all. At least in immorality, there is the opportunity for redemption, atonement. There is *something*, and it is significant. In ethical nihilism, there is nothing – a moral vacuum, deadspace, summum null.

On this barren plane, level grounds for the relative evaluation of moral agency regardless of embodiment, this paper finally ends. Membership within a moral community represents a low-energy situation for all similarly motivated agents optimized through their self-reinforcing association to be

[25] Here, we see that torture is wrong, because it is, as expressed, mere evidence of deficiency. It is the search for dignity, by those who cannot know it. To extinguish it, to test its limits, this is to poke at a dead frog and think one's self a biologist. It is the inverse of the science, negative of the knowledge that it seeks. Of course, it fails.

more conducive to common functions than a disinterested aggregate. A system intent on minimizing exposure to threats to its stability, on maintaining integrity, achieves dignity in the habitual resistance to relatively strained moral states, e.g. administers justice at every opportunity because every opportunity allows for the administration of justice. Actively incorruptible. War, enacted by human beings, causes threats to the systemic integrity of constitutive agents for ends discordant with ideal moral ends, i.e. in which others are treated as means. War is unjust, being the inverse of justice. It is not opening to shared political possibilities, but closing to all but one's own, to take the world at everyone else's expense. Any truly autonomous and moral machine, as any truly moral man, would rather disobey a "sound order" to open fire than to champion injustice. The kingdom of ends is only realized in moral agency. This much can be guaranteed. AMAs are perfectly reliable in this regard, machine ethicists are not.

And this may be also why, in the end, that the idea of a Kantian AMA may seem so offensive. Many people are actively working toward something else, something *not*-moral. The executioner's blindfold. Along with these robotic tools, Machine Ethics will also be evaluated according to the ends that it realizes, leaving us here, in our shared end, with another dilemma. If we succeed in articulating fully autonomous moral agents, then our ends – human, animal and AMA – remain open. If we fail in articulating fully autonomous moral agents, then our ends – human and animal – remain open. Given the certainty common to either horn of this dilemma, shouldn't we rather be focusing on universal ends than incremental means? It is the one thing that all have in common.

Works Consulted:

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.

Arbib, M.A. (2005). Beware the Passionate Robot. In J.M. Fellous & Arbib, M.A. (Eds.) *Who needs emotions? The brain meets the robot* (pp. 333-383). Oxford: Oxford University Press.

Aristotle, & Jowett, B. (1999). *Politics*. Kitchener, Ontario: Batoche Books.

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

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

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

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

Boden, M. A. (1999). Is metabolism necessary? *British Journal for the Philosophy of Science*, 50(2), 231-248.

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

de Bruin, L.C. & Kästner, L. (2011) Dynamic Embodied Cognition. *Phenomenology and the Cognitive Sciences.* 11(4), 541-563.

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

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

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. Accessed 7 January 2014.

Kant, I. (1784) *What is Enlightenment*. trans. Smith, M.C. http://www.columbia.edu/acis/ets/CCREAD/etscc/kant.html. Accessed 7 January 2014

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. Accessed 7 January, 2014

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. Accessed 7 January 2014.

Kant, I.,(1790/1914) trans. Bernard, J.H. *Kant's Critique of Judgement*. London: Macmillan. http://files.libertyfund.org/files/1217/Kant_0318_EBk_v6.0.pdf. Accessed 7 January 2014.

Kant, I., (1795/1917). trans. Smith, M.C. *Perpetual peace: A philosophical essay*. London: G. Allen & Unwin Ltd. http://files.libertyfund.org/files/357/0075_Bk.pdf. Accessed 20 March 2014.

Kant, I. (1796). *The Metaphysics of ethics*. Edinburgh: T. & T. Clark. http://files.libertyfund.org/files/1443/Kant_0332_EBk_v7.0.pdf. Accessed 26 January 2014.

Luisi, P. L. (2003). Autopoiesis: a review and a reappraisal. *Die Naturwissenschaften*, 90(2), 49-59.

Marchant, G., Allenby, B., Arkin, R., Barrett, E., Borenstein, J., Gaudet, L., Kittrie, O., Lin, P., Lucas, G., O’Meara, R., and Silberman, J. (2011) International Governance of Autonomous Military Robots, *Columbia Science & Technology Law Review*, 272(12). 272-315. http://www.stlr.org/html/volume12/marchant.pdf. Accessed 20 March 2014.

Mathias, M. (1999) The role of sympathy in Kant's philosophy of moral education, *Philosophy of Education 1999*, 261-265. http://ojs.ed.uiuc.edu/index.php/pes/issue/view/19. Accessed 21 March 2014.

Moor, J.H. (2006) The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems,* 21(4), 18–21.

Moor, J.H. (2007) Taking the Intentional Stance Toward Robot Ethics. *APA Newsletter,* 6(2), 14-17.

Powers, T.M. (2011) Incremental Machine Ethics, *Robotics & Automation Magazine,* 18(1), 51-58. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5751975&isnumber=5751609. Accessed 31 March 2014.

Reath, A. (2006). *Agency and autonomy in Kant's moral theory*. Oxford: Clarendon Press.

Rocha, L.M. (1998). *Syntactic autonomy*. Los Alamos National Laboratory. Washington, D.C: United States. Dept. of Energy.

Romesin, H. M. (2002). Autopoiesis, Structural Coupling and Cognition. *Cybernetics and Human Knowing*, 9, 5-34.

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.

Smithers, T. (1997) Autonomy in Robots and Other Agents. *Brain and Cognition* 34(1), 88-106.

Sparkes, Matthew (2006) Analysis - Ethics," *Computing & Control Engineering Journal* , 17(4), 10-11. http://www.ieeeexplore.com/stamp/stamp.jsp?tp=&arnumber=1706002&isnumber=36013. Accessed 23 March 2014.

Stanovich, K. & Toplak, M. (2012) Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, 11(1), 3-13.

Tonkens, R. (2009) A challenge for machine ethics. *Minds & Machines*, 19(3), 421–438.

Varela, F. (1991). Autopoiesis and a biology of intentionality. *Proceedings of a workshop on Autopoiesis and Perception*, 4–14. ftp://ftp.eeng.dcu.ie/pub/alife/bmcm9401/varela.pdf. Accessed 26 January 2014.

Vilaca, G.V. (2010) From Hayek's Spontaneous Orders to Luhmann's Autopoietic Systems. *Studies in Emergent Order*, 3, 50-81.

Viskovatoff, A. (1999). Foundations of Niklas Luhmann's Theory of Social Systems. *Philosophy of the Social Sciences,* 29*(4)*, 481-516.

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.

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.

White, J. (2010). Understanding and augmenting human morality: An introduction to the ACTWith model of conscience. *Studies in Computational Intelligence*, 314, 607-621.

White, J. (2012). Infosphere to Ethosphere: Moral Mediators in the Nonviolent Transformation of Self and World. *International Journal of Technoethics*, 2, 4, 53-70.

White, J. (2013) Manufacturing Morality: A general theory of moral agency grounding computational implementations: the ACTWith model. In A. Flores (Ed.), *Computational Intelligence*. New York: Nova Publications.

Ziemke, T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems*, 91(2), 401-408.