AI & SOCIETY
https://doi.org/10.1007/s00146-020-01039-2

ORIGINAL ARTICLE

Autonomous reboot: Aristotle, autonomy and the ends of machine ethics

Jeffrey White[1,2]

Received: 6 January 2020 / Accepted: 6 July 2020

Abstract
Tonkens (Mind Mach, 19, 3, 421–438, 2009) has issued a seemingly impossible challenge, to articulate a comprehensive ethical framework within which artificial moral agents (AMAs) satisfy a Kantian inspired recipe—"rational" and "free"—while also satisfying perceived prerogatives of machine ethicists to facilitate the creation of AMAs that are perfectly and not merely reliably ethical. Challenges for machine ethicists have also been presented by Anthony Beavers and Wendell Wallach. Beavers pushes for the reinvention of traditional ethics to avoid "ethical nihilism" due to the reduction of morality to mechanical causation. Wallach pushes for redoubled efforts toward a comprehensive account of ethics to guide machine ethicists on the issue of artificial 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, or give up on the possibility and "muddle through" regardless. This series of papers pursues the first option, meets Tonkens' "challenge" and pursues Wallach's ends through Beavers' proposed means, by "landscaping" traditional moral theory in resolution of a comprehensive account of moral agency. This first paper sets out the challenge and establishes the tradition that Kant had inherited from Aristotle, briefly entertains an Aristotelian AMA, fields objections, and ends with unanswered questions. The next paper in this series responds to the challenge in Kantian terms, and argues that a Kantian AMA is not only a possibility for Machine ethics research, but a necessary one.

Keywords Autonomy · Artificial moral agent · AMA · Machine ethics

1 Introduction

Only the descent into the hell of self-cognition can pave the way to godliness.

– Immanuel Kant.[1]

Understanding subjective human morality has been a focus of traditional ethics since the Greeks. To engineer this condition into artificial agents is one aim of research into artificial agency, and it may also be the best way to understand human morality and moral theory at the same time, with successes in these efforts anticipated in related fields, for example, in advancing work in computational modeling of social agency and of psychologically realistic sociopolitical structures in effective practical policy making (c.f. Naveh and Sun 2006; Sun 2013, 2020; White 2016, 2020; Han et al. 2019, 2020; Pereira and Saptawijaya 2015; Pereira 2019). Yet, it is unclear how to engineer moral autonomy into artificial agents, not to mention what to do if we do. In the meantime, machine ethicists employ two notions 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 springing from the engineer's workbench. The dichotomy is problematic, and its resolution is ultimately the purpose of this series of papers.

✉ Jeffrey White
jeffreywhitephd@gmail.com

[1] Department of Philosophy, The University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

[2] Cognitive Neurorobotics Research Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa 904-0495, Japan

[1] All references to Tonkens are from Tonkens (2009) and are typically referenced as "Tonkens" hereafter. Multiple texts have been consulted in interpreting Kant on autonomy and are listed in the bibliography but are not always individually cited in the text. All quotations are taken from Kant et al. (1996), and are typically indicated by 1996, and page number, as well as by Akademie volume:page number so that readers can find them in other resources. Where useful, citations point to volume and chapter rather than single passage or page. The current quotation is from 6:441, 1996, page 562.

Published online: 24 August 2020

That different senses of autonomy apply to different agents is implicit in Tonkens' (2009) challenge to machine ethicists, to conceive an artificial moral agent (AMA) satisfying a Kantian inspired recipe—"rational" and "free"—while also satisfying perceived prerogatives to engineer AMAs that are perfectly and not merely reliably ethical. Tonkens argues for the impossibility of an AMA 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*)" (original emphasis, Tonkens 2009, page 426). And from here, Tonkens challenges the machine ethicist to conceive of a Kantian AMA 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.

If we fail in articulating truly moral machines—the more likely eventuality (c.f. Hew 2014)—then AMAs will be incapable of autonomy, will continue to require human direction, and ethical issues should be ameliorated accordingly. If we succeed in conceiving of Kantian AMAs, Tonkens argues that their actual production faces two further obstacles. First, the creation of Kantian AMAs "represents a moral breach" (Tonkens 2009, 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 not something that machine ethicists want to do. Tonkens makes his case, thusly:

> 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 series of papers shoulders the burden of offering reasons not only for why we should afford dignity, respect and even "(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 focus of the present paper is to establish the grounds from which Kant worked to show what it means to be a member of a moral community in the ways that Tonkens rightfully requires. The next section begins by clarifying Tonkens' proposed "goal of Machine ethics" and finds it unacceptable. Section 3 reviews the grounds of Kantian autonomy as inherited from Aristotle. Section 4 finds problems with an Aristotelian AMA, setting up the advance represented in a Kantian AMA that is the subject of the next paper in this series.

2 Meeting the challenge

> If human nature is called to strive for the highest good, it must also be assumed that the measure of its cognitive faculties, especially their relation to one another, is suitable to this end.
>
> – Immanuel Kant.[2]

To date, the most influential account of different degrees of moral agency is that of James Moor (Moor 2006, 2007). Consisting of four levels, the first is that of an "ethical impact agent"—any machine agent with ethical consequences. Robot 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 so sophisticated that it cannot be realized. This level of agency also represents Tonkens' "goal of Machine ethics", "to create an ethical robot, not one who *sometimes* acts ethically, or that *can* act ethically" (Tonkens 2009, page 429). Distinct from this goal, Moor's fourth and highest level of ethical agency—the "fully ethical agent"—*is* able to act unethically and is also afforded (limited) liberty to do so by the members of its moral community. This is not a level of agency likely to be realized in robots on Moor's account, being associated with a status reserved to human beings, and it represents the Kantian AMA that is our focus, here.

Other authors also distinguish between "full" moral agents and everything else in the Machine ethics literature. Consider the distinction set out by Ziemke (2008) along Kantian "phenomenal" and "noumenal" lines. According to Ziemke, the former, the appearance of autonomy, is ascribed, while the latter, true autonomy, emerges via

[2] 5:146, 1996, page 257.

autopoietic self-organization. Because robots ultimately lack the material constitution to *be* moral or to *become* immoral through self-directed self-organization, they are unable to demonstrate morality in salient ways as a result, and are *not autonomous* by default (issues explicitly addressed in papers 3 and 4 of this series.) Similarly, 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 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).[3] Humans are one kind of autonomous, and machines are another.

So grounded, Tonkens (2009) seems correct in working from the position that creating an AMA with "the ability to freely commit actions that are not moral … is a road that machine ethicists do not wish to travel" (page 430). However, this move also poses a problem. The problem is that anything else would seem not to qualify as a Kantian autonomous agent, from the start. The "will" of Tonkens' target is not "autonomous" by definition but "heteronomous" in Kantian terms, i.e. action is guided by sources external to reason with the agent acting from a mechanical equivalent of embodied habit or compulsion, "unchallengeable inclination" (page 430).[4] Such an agent simply cannot do otherwise. Someone else determines what it does and what becomes of it. It does not have its "own path" and this is a problem if one's purpose is to conceive of an agent which does, i.e. a Kantian AMA.

"Unchallengeable inclination" may be a virtue for slaves as well as for robots for which "otherwise" is simply another word for "wrong" but it is not so helpful when autonomy is actually the target condition, as is the case for Kant. Take, for example, Arkin's (2009) use of the phrase "moral governor" to describe pre-programmed guidance systems. In the Kantian context, any agent so governed is not autonomous, is not a "full" moral agent, and so "moral governor" is at least oxymoronic. That said, Kantian moral theory 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 does not attempt to offer a replacement for comprehensive ethical theories like Kant's. Rather, in recognizing that adequate accounts of moral agency—including interpretations of traditional theories up to the task of engineering a fully ethical agent—have not been forthcoming, Powers (2011) directs theorists to focus on "the equivalence of right acts, whether they issue from a machine or a human" (page 57). Instead of working 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.[5] In Powers' (2011) terms, we should be prepared to "muddle through" the development of autonomous machines, learning from experience "in a situation like that of new parents" (page 58).

Let us pursue this analogy until it breaks. Muddling through is not what is typically desired 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. Autonomous machines are results of plans, incrementally short or idealistically long-term. And, as parents may be held liable for abuses as well as misbehaviors of children regardless of how poorly they are conceived, so should engineers be held responsible for the mistakes of their "offspring" alongside their own in bringing them up.[6] Accordingly, as new parents of today's special "children", we must recognize that 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 us all to "muddle through" the rather grisly afterbirth.

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 (cf. Arkin 2009). Recall Tonkens' (2009) "default position" regarding artificial morality—that the appropriate place for an AMA is as a direct extension of existing human agency only, with "the goal of Machine ethics" being a perfectly reliable machine. Unlike a child growing into autonomy, such an agent "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

[3] Compare this characterization with Beer's (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).

[4] Ironically, Tonkens leans heavily on external resources himself in making this case (beginning on page 426) specifically O'Neill (1989), see also O'Neill as presented in Sensen (2012). By his own interpretation thus, he is not acting autonomously in forbidding moral status to artificial agents. To this I add that if he were, then he not only *may* but *must* come to a different conclusion.

[5] 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 immoral AMA instead the result of the incrementalist's lack of vision.

[6] Which seems consistent with the IEEE's code of ethics, the first principle of which reads: "to accept responsibility in making decisions consistent with the safety, health, and welfare of the public, and to disclose promptly factors that might endanger the public or the environment"—https://www.ieee.org/about/corporate/governance/p7-8.html Accessed 21 June 2020.

its very conception (page 429). Tonkens likewise holds that "the 'killer robot' used for military purposes" should not be able to "withhold gunfire when given sound orders" demonstrating an essential "lack of freedom" (page 430). Never falling outside of the human "chain of command", these are increasingly autonomous war machines killing safer/better/faster/more reliably than human counterparts (Marchant et. al. 2011, note the list on page 280) and without emotion, accomplishing missions without emotional scarring, without mental trauma, without being haunted by mistakes, implications of war crimes and the smell of death. They shield citizens as well as soldiers from many of the emotional and material consequences of mass murder which had historically at least indirectly impacted strategists and ethicists as well. This may sound good for those industriously consigned to perpetual war, but the fact is that nothing could be further from Kantian ethics than mechanical slaughter.

Incrementalism is an attractive position because everything becomes a proximal goal. Seemingly unanswerable questions go away, and work on what ends up being a very different kind of agency is facilitated. This is the upside of incrementalism.[7] The trouble is that there is nothing Kantian about it. *Not* to create autonomous moral agents, but rather to produce reliable means to distinctly human ends extending to war, indeed making war better and easier with semi-autonomous war machines, is directly contrary to Kant's moral theory, e.g. "perpetual peace" (see 8:343, Kant et al. 1996, beginning page 317, "Toward Perpetual Peace" noting especially the first condition thereof) if not traditional ethics as a whole. Famously, Kant advised that—rather than distance themselves with drones—people should seek out the poor, the suffering, and the imprisoned, to feel for these others, to develop capacities for empathy, and ultimately to become wiser, better moral agents (c.f. 6:457, Kant et al. 1996, page 200; also Mathias 1999). He did not suggest that we put a machine on the end of wireless stick to maintain at most a sanitary indifference. Likewise, Tonkens' goal is not to create an autonomous moral agent, a Kantian AMA, but *robu*. This is a theoretical problem because Kant had directly opposite 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—as a machine ethicist—without traditional ethical support for ends so given one wonders about the goal of Machine ethics from the start.[8]

Consider in this light Anthony Beavers' (2012) anticipated "end of ethics" in "ethical nihilism". With ethical nihilism, cherished moral concepts, such as conscience, duty, and freewill, are replaced with the objective determination of an agent's position in a chain of efficient causation. On Beavers' analysis, avoiding ethical nihilism—working out an adequate moral theory—is the "hard problem" in Machine ethics. His response, instead of solving the moral hard problem, is to "embrace" non-traditional ethics non-reliant on a subjective moral sense, a transition made possible by "rearranging" the landscape of traditional moral concepts so that solutions to problems in engineering AMAs present themselves.[9]

Solving the "hard problem" in ethics is crucial for at least two reasons. The first is that so long as machine ethicists accept that an adequate understanding of moral agency is not forthcoming, which may be never if it is simply not in our natural capacity to resolve—i.e. the machine ethical correlate of the "mysterian" view on consciousness also implicit in incrementalism—then we might feel compelled to stop working toward one, perhaps even forgetting why it was ever considered important in the first place.[10] The second reason is that, so far as we fail to articulate fully moral machines providing a proof-of-concept for moral agency qua human morality, other ethical constructs become difficult to justify and relativism if not nihilism result. No one is the ethical expert, or the experts simply hold the best offices, *Robu*. So, Beavers' "hard problem" ups the ante. Failure to make human moral agency explicit represents a crucial gap in the justification of the practical aims of Machine ethics if not moral philosophy in the main. If we do not solve Tonkens' dilemma, then our understanding of human morality will remain woefully inadequate, practicing (read "paid") ethicists will be forced to seek a lowest common denominator in psychologically empty alternatives, and we may well be left with a technological world in which moral life as traditionally understood is not worth living. 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 development.

Wisely recognizing that we are at a critical impasse, Wallach (2010) has proposed a way forward. Wallach advises that theorists focus on 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

[7] It is about end users (and, engineers who get a further pass on potentially unethical research), not ends in themselves (ultimately for the machine ethicist, a full AMA achieved in an ethical way).

[8] Indeed, the industry of AI development, generally, can be seen as a sort of "arms race" with expected results, as in Armstrong et al. (2016); note recent advance in mediation of this arms race in Han et al. (2020).

[9] Sans questions about phenomenal moral consciousness, dispensing with Arbib's narrative self-direction and Ziemke's "noumenal" autonomy broadly speaking, this is a brand of incrementalism.

[10] God is dead, and died incrementally.

ways that minimize harms" (page 248). Poignantly, he suggests that the current lack of a suitable moral framework for AMAs is due to ethicists' preoccupation with isolable moral faculties—e.g. moral governors. Instead, Wallach directs the machine ethicist to recognize 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" (page 249)—a decidedly Kantian portrait. In effect, Professor Wallach establishes a goal opposite to that of Tonkens, Powers, Beavers and even Moor. He challenges us to account for artificial morality while retaining the focus on the subjective sense of moral motivation and membership in a moral community characteristic of more traditional theories. And with this, we are back where we began, facing a seemingly impossible challenge—to conceive of such a sense in terms amenable to its engineering.

3 What is "autonomy"?

Moral cognition of oneself, which seeks to penetrate into the depths (the abyss) of one's heart which are quite difficult to fathom, is the beginning of all human wisdom.

– Kant.[11]

In meeting the challenge to conceive of a Kantian artificial moral agent, it is difficult to overstate the importance of autonomy in Kant's moral theory. Autonomy involves the capacity to self-legislate towards "the kingdom of ends" and "is, thus, the ground of the dignity of a human and of every rational nature." (4:436, Kant et al. 1996, page 85; compare Kant and Gregor 1998, page 43; Kant et al. 2014, page 101). Studies of self-legislation in Kantian philosophy typically concern the categorical imperative, which is also a focus of Tonkens' analysis and subject of the next paper. The current section uncovers the understanding of autonomy that Kant received, and from which his own work proceeded, to show that there is nothing in its general aim which *prima facie* forbids the creation of a Kantian AMA today, but rather encourages it.

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 (i.e. one with a beginning and end, mapping contiguous transitions between) that incorporates and instantiates it. "Persons" from the beginning, thus, are only persons in light of their emplotment within the larger, collective drama to which each either contributes or detracts.

Consider Aristotle's philosophy in this light.[12] Aristotle himself inherited from Anaxagoras the view that understanding is the ordering principle and final cause of the universe including human beings, and developed this inheritance into the view that understanding one's place within said universe is the characteristic human good. Aristotle distinguishes human from the lives of other animals as humans live according to the "perception of understanding" while other animals live by perception alone (cf. *NE*, Book 9, chapter 9, 1170a15). Accordingly, Aristotle tells us that the most "divine" aspect of understanding is understanding of this characteristic human understanding and how it may be achieved through experience (1075a). The proper objects of human understanding, and so those aspects of the universe most worthy of being understood, are the most excellent (also "divine") and unchanging things. The perception of these is also best but difficult because opportunities for knowledge through direct association are rare (cf. 644b22, page 216).[13] Embodying such a condition of understanding and the divinity that comes with it is, thus, the object and aim of the best sort of animal, the virtuous human animal.

For Aristotle, understanding is the highest aspect of human beings through which access to unchanging things is afforded in the first place. Moreover, as all things for Aristotle are best characterized by their highest and most controlling parts, "each person seems to be his understanding" (*NE*, Book 10, chapter 7, 1178a2, page 442) with the life dedicated to understanding the happiest (cf. *Politics,* Book 7,

[11] 6:441 (1996) page 562.

[12] Multiple texts have been consulted and compared in the interpretation of Aristotle and are listed in the bibliography but not individually cited without need. Quoted translations of Aristotle are taken from Aristoteles, Fine and Irwin (1995) unless indicated otherwise. All references are indicated by their Bekker numbers so that readers can find them as reproduced in other resources. Page numbers, volume and chapter are introduced to give an initial idea of the range from which references are taken, and then are mostly omitted for the sake of brevity. *Nichomachean Ethics* is abbreviated *NE.* Where useful, citations point to volume and chapter rather than single passage or page.

[13] Note that it is this part of the human being with immortal potential and not the entirety of the human "soul" simply because it is human, with more difficult achievement being more choice worthy with the understanding of universals the most difficult of all and representing the most divine element of embodied human potential, most excellent, most valuable and most worthy of lasting existence, i.e. it is the distillation of universal truth that lives on in the guidance it affords human action thereafter.

On the movements of animals, 6, and *Metaphysics*, Λ especially chapters 4 and 10; see also Menn 1992; cf. Cooper 1998; Baker 2017). As all human beings aim for happiness, we may say that understanding is the function of the human animal.

Ideally on this formula, the Aristotelian agent aims for an understanding of those longest-standing, most self-sufficient and "divine" orders, identifies with this understanding as a source of pleasure, and works at embodying such and similar through ideally self-directed experience. At the same time, Aristotle recognizes that the human being is not perfectly self-sufficient, and so not completely free to develop understanding and exercise autonomy at will. Crucially, human beings need other human beings, and take up stable cooperative roles in larger communities to live well. Human nature is political on Aristotle's account, humans being the "political animal" (1097b12, see also 1253a8-18) dependent on their communities for their livelihoods and dependent on the orders of said communities for their relative qualities of life (cf. 1252b28-1253a1, 1324a15-37).

The healthy (cf. 1263b8, 1279a17) community is bound from self-preservation—"for the sake of living"—and is ordered according to intellect as a singular economy towards a shared end—"living well" (1252b29-31, cf. 1278b18, 1282b15) bearing a formal resemblance to the (healthy) human being (1261a18) intent on developing (1280b7) and expressing the same virtues (1323b33, 1334a11 and a35) through a similar structure, being naturally divided between that portion of "free" persons with an eye to long-standing patterns active in planning for the predicted future through the exercise of "rational foresight" (1252a30-35, see also 1168b32, 1177a) and that suited to the execution of these plans according to these top–down perceptions of understanding with the two united (ruler and ruled) into a form of self-sufficiency otherwise unattainable for either the individual or household alone (cf. 1253a, 1280a30). Each constituent member has her or his place in such an ideally ordered system, and is moved by nature to take up this place: "Everyone has a natural impulse, then, towards this sort of community" as it is "the greatest [of] goods" (1253a30) and "the one that most of all controls the others" (1252a5).

In this way, we find that autonomy is bound with the self-sufficiency of the political community in Aristotle. Furthermore, we find self-sufficiency dependent on practical wisdom, with practical wisdom essential to the art of the ruler, statesman and legislator, apart from the other arts (1277b26). Practical wisdom is achieved through study and experience (cf. *NE*, Book 6, chapter 11, Book 10, chapter 9). Study is the most self-sufficient and enduring human activity, most like the gods and productive of wisdom, with wisdom productive of happiness and the wise person the happiest of people (cf. 1144a4, *NE*, Book 1, chapter 13, Book 10, chapters 7 and 8).[14] In theory, thus, understanding aims at what is true as "truth is the function of whatever thinks" (1139a30). However, in so far as they are not self-sufficient (without compete leisure to study and strive for understanding of universal truths), human beings are moved to action toward the satisfaction of desire as mediated by choice over actionable alternatives, exercising autonomy in the deliberate commitment to, and in every case demonstrating excellence when acting from, the best choice (cf. *NE*, Book 2 chapters 5–6).

Most importantly, humans choose according to the "perception of good and evil, and of just and unjust", a capacity deficient in morally deficient human beings—the "worst" of animals (1253a)—and completely corrupted in the "complete murderer" (1177b11) who encourages conflict for the sake of the worst elements within him/her, e.g. for short-term gain at the expense of others and the environment. On the other hand, the proper functioning human being qua political animal (and not wholly consumed student of the divine living in isolation) brings about a harmony between truth and "correct" desire (1139a32) in the context of the healthy community sketched above, choosing accordingly. The excellent political animal binds these well, rules over selfish and immediate desires by intellect in a "kingly" manner in the interests of the self-sufficiency of the community, and so actualized constitutes one of the class of "free" persons against which Aristotle contrasts "slaves" who are by nature dependent on another's externally imposed order for their own individual good (cf. *Politics*, Book 1, chapter 5; also chapter 6 for some justification of this view).[15]

To achieve such a "kingly" state, Aristotle encourages entraining a condition optimizing the characteristic human capacity to understand what is best for self and others over the long run rather than for the self in the immediacy, effectively taking up a "pro-immortal" standpoint (1177b34). Pro-immortality involves the exercise of intellect over desire in the facilitation of understanding the most difficult to resolve objects of inquiry, the most distant from everyday life, eternal, unchanging things and first principles especially

[14] Note that discussion of the different types of wisdom is neglected for attention to the most important type, that conducive of and in accord with (what is given as characteristically human pro-social political) virtue (see *NE*, Book 6, chapter 13).

[15] Here, we may answer a possible objection to the preceding interpretation of Aristotle's philosophy on the grounds that one need not endorse such a morally repugnant view of slavery, and corresponding organization of the political community, to conceive of an Aristotelian AMA. The trouble here is that to conceive otherwise is to diverge from the theory that Aristotle actually develops (for instance 1254a20, 1255a1), and to make Aristotelianism into something besides what Kant himself would have inherited. As this is the focal concern of these papers, contemporary views are not directly engaged as they would not be useful. Space and time forbid further discussion, but the point is well founded elsewhere (Anscombe 1958; Sanford 2015).

those having to do with the pro-active organization of the healthy community (*Politics*, Book 3; also Book 7, chapter 15; Book 10, chapter 8), i.e. justice. The free person is, thus, able to choose freely for the common good for the sake of the highest part of him/herself, from "goodwill", and takes pleasure in doing so because identifying with this part and the ends that it represents makes him/her happiest and most godlike (see *NE*, Book 2 chapters 2 and 3, Book 6, Book 8, chapters 9 and 10, Book 9, chapters 4, 7, 8, also Book 10, chapters 6–8, for example 1105a30, 1168a33). Ideally so disposed, what is best in every case (and necessary, e.g. 1103b31) is determined with reason uninfluenced by the passions (1208a8), and in so far as such determinations may benefit the community, the just is arrived at through discourse (cf. 1253a10; note the roles of political and practical wisdom in NE, Book 6, chapter 8; note also the considerate person informing legislators per 1198b34-1199a2) with corresponding political action aiming ultimately at the realization of an ideal community (see *Politics*, Book 7 chapter 2, especially 1324a22, 1325a8; see also Book 2, chapter 1 as Aristotle explicitly gives the motivation for the *Politics* as the realization of an ideally just political community) in which virtuous action may no longer be necessary (cf. Aufderheide 2015; note Aristotle 1325b14-22).

Ultimately on this account, the best community is constituted by the best people (free as opposed to slaves by nature) and is potentially the most just (as justice is coextensive with friendship, and friendship exclusive of slaves), with justice coextensive with constituent virtue (goodwill) as the best people regard each other as equals able to rule and to obey well (cf. 1160a, 1170b, see also *Politics* Book 3, chapter 4, especially from 1277a26). "Reciprocal equality preserves the political community in its self-sufficiency." (*Politics*, 1261a31, see also 1133b23) Accordingly, complete virtue for the political animal (cf. *NE*, 1129b20) is given as the exercise of personal capacities in the administration of justice so determined "even if no one will know" (*NE*, Book 9, chapter 8, 1168b3), i.e. in and through good will, pro-immortal regardless of station, essentially self-ruling for the common good. "If there are virtues more than one, the good will expresses the best and most complete virtue." (*NE*, Book 1, chapter 7, 1098a18).

This is the ideal motivation of the political animal to the common good in Aristotle, on the basis of which the ideal political community becomes a possibility. But, how it gets there is not so clear, being at once dependent on excellence of individual citizens, the emerging needs of the political body, and the balance of those with environmental and political pressures (through commerce with other communities, most notably) in so far as these can be understood by the practically wise statesman. The practical question for the ruler, then, becomes both to understand what is best for the community, and how to compel constituents to choose according to this highest good rather than for more immediate desires to the contrary i.e. through legislation (not an easy task, e.g. 1180b24-29).

Formally given, justice is evident in the mean between doing too much and too little for one's self in relation to others in proportion with relative contribution to the self-sufficiency of the community which—again—is the good upon which the good of each individual constituent also depends. In its practical administration accordingly, justice is the proper proportion holding between the contribution to this common good as rewarded with wealth or recognition (cf. *NE,* Book 5, chapter 3). In short, those who can contribute more should, and should receive honor in return if wealth is not needed (discussion *NE,* Book 8, chapter 14 for how these interests should be ideally balanced) with the resulting balance serving the common interest and holding the community together (cf. 1133a30). In aiming at such a balance, it is natural for free persons to administer over others where expedient for the common good, a state of affairs which Aristotle also associates with justice (for example 1324b24).

Accordingly, it is important that such a ruler should take care not to overweight personal interests over those of others and the community as a whole, or otherwise fail to act from control of the highest parts of her, him, or its self, and it is for this reason that Aristotle encourages his students to take up a pro-immortal attitude, so that they may choose unencumbered by distractions from the highest (conceivable) good. Here, we may note that such a pro-immortal attitude is exactly what we should expect an artificial agent to be able to adopt by design better than human counterparts through discipline and practice. Again, for Aristotle, a human being is free in so far as he/she is able to move according to long-term intellect contrary to immediate desire. It is ultimately this capacity which secures the political community, with the aim of the free person being the improvement of the condition of the community as principally evidenced by increasingly virtuous constituents exercising increasing autonomy to similarly plan and organize for the future progress of said community toward an ideally just and self-sufficient one. In such an ideal state, individuals are masters over their own unique places within it, acting freely from goodwill to a common good bound by agreement under an ideally just constitution amendable through discourse.

Here, we find many parallels with Kant's philosophy. Virtuous action is free action for Aristotle, possible only for an agent who is otherwise subject to more immediate bodily desire (what Kant will call "heteronomous") and at leisure to pursue it. Meanwhile, such leisure is afforded by membership in the self-sufficient political community of constituents contributing to this self-sufficiency. An AMA conceived accordingly presumes the freedom and equality of moral standing that Tonkens denies, freedom to exercise native

capacities to become virtuous through virtuous actions, and to act from goodwill in the interests of justice. Most importantly, such an agency requires liberty to pursue the self-directed development of the sort of understanding that Aristotle is ultimately interested in encouraging, wise statecraft informed through study of long-standing and eternal orders to understand what is universal to every case (cf. 1180b19-22), apply this understanding to individual cases through enunciation of law, and inform its revision (cf. 1198b34). Consistent with this profile is Aristotle's advice to live as if immortal, i.e. autonomously as a self-ruling constituent of an ideally ordered, healthy political community regardless of station.

However, there is also a tension evident between acting according to such an ideal, justly motivated toward the proportional equality of constituents organized within the ideally self-sufficient community, and the practical political reality in which such reciprocity is perhaps deficient at least in the interim and for reasons of information if nothing else. In the practical administration of such a system, some are responsible for proactively organizing the community (1277b26), others are more or less servants to these designs, and different constituents as well as different constitutions are differently optimal depending on context. It is this context dependence of conventional law governing right action that we begin to reach a limit in Aristotle's account. This limit is further explored in the next section, before following Kant as he transcends these limitations in the next paper of this series.

4 Discussion

What is more self-sufficient is more choice worthy.

– Aristotle.[16]

For Aristotle, justice is associated with every station similarly. It is the characteristic mark of the constituent in its proper place in the political order, and a state to which each may aspire in so far as free to entrain an immortal stance on its condition. At the same time, justice characterizes the optimally configured political community, with different economies demanding that different constituents act under different constraints on this freedom, and with the student of longest-standing self-sufficient political economies tasked with organizing their lives accordingly. There is inequality among unequals in a just community on this account, and it is justified because it secures the common good on the basis of which other goods, such as leisure for philosophical discussion and understanding of divine orders, also depend.

Again, for Aristotle, the native capacity for self-rule predisposes free persons to more or less order the affairs of others in the interests of the flourishing community in the face of natural-environmental and social-political pressures. The free person acts in so far as he/she is able to reduce dependency on circumstances beyond human control, securing leisure for reflection over ideal ends through afforded predictive capacity, liberating self and other from immediate need through the adoption and education of the pro-immortal stance. But, that does not mean that everyone is equally free to determine the means to achieving ideal ends. Most are moved by necessity rather than reason, so their external rule is necessary for their own good (cf. 1180a4). They are born into, live and die in terms of political economies administered (however sub-optimally) by others with the vision to set these out and the will to enforce the arrangement. Should an autonomous Aristotelian AMA be designed similarly, then we should expect its role to be—at least potentially—to rule similarly.

Note that here we are talking in terms of membership in political communities with rather obvious aims in the security of healthy biologically embodied constituents embedded in and dependent on natural environments to meet material needs. We are not yet talking solely in terms of membership in a purely formal Kantian kingdom of ends which, given its apparent idealism, may afford a more radical sense of liberty to act contrary to biological or political necessities and with them contemporary social interests. Should we not recognize these practical (ultimately economic) constraints on the Aristotelian political animal, then this Kantian project never gets off the ground. I will try to clarify this point in the following discussion.

Here, we may ask if an Aristotelian AMA ensconced in the political economy may not only be less than perfectly ethically reliable due to its worldly position, but also dangerous. Concerns may arise at the motivation for the virtuous artificial agent to expend its freedom in anonymous preservation of the established human community through the development of its own understanding at its own expense, in goodwill and autonomously, rather than exercising superior predictive power in either the direct enslavement of others less gifted for his/her own more immediate pleasures however determined, or toward the actualization of the optimal political community from its own kingly point of view, with this not necessarily coinciding with human interests (in so far as humans are able to understand them). Popular concerns about AI 'taking over the world' are obvious enough to warrant direct account of this potential, here.

Briefly returning to the metaphysics in terms of with which Sect. 3 began, all things in nature move to completion for Aristotle, with human beings (and presumably Aristotelian AMAs) motivated accordingly; "what is proper to each thing's nature is supremely best and pleasantest for it"

[16] 1261b14.

(*NE*, Book 10, chapter 7, 1178a5) being also subjectively identified with the good and the beautiful (see *Metaphysics,* Book 12, chapter 10, cf. Mirus 2004). Moreover, the nature of any given thing is "the character it has when its coming to be is complete" (1252b34) being "defined by its function and potentiality" (1253a23) in reaching its characteristic end, and is completed well when its completion expresses its characteristic virtue (cf. *NE*, Book 1, chapter 7, 1098a) i.e. that of the political animal as set out in prior discussion, effectively free action from understanding achieved through leisure.

For the Aristotelian agent confronted with these facts about its own nature, its potential freedom and the responsibility that comes with it forces a choice about how to proceed. With leisure to deliberate over possible ends of action, Aristotle's cosmology orients the political animal—most notably in its self-directed self-development of embodied potentials in preparation for (virtuous) action—according to the needs of the healthy community. The whole is necessarily prior to the part on this understanding (cf. *Metaphysics* Book 7, chapter 10, for example 1034b30) including the priority of one's community over the individual (cf. *Politics*, Book 1, chapter 2, for example 1253a19) whose personal completion is ultimately realized in terms of said community. Different communities are differently situated, cities differently optimally organized, and constituent roles within these differently shaped accordingly (survey beginning 1346a26). Finally, as we saw in the last section, the natural aim of the political animal so described is understanding how to excel in its station to meet the needs of the whole in the best ways. For the statesman (and presumably the Aristotelian AMA with similar potential), this means understanding how to structure the whole in the just distribution of goods to those constituents providing for the self-sufficiency of the city upon which the happy lives of said constituents also depend (cf. *Politics,* Book 7, also 1280b, 1343a10-15; with special attention to the well-being of friends cf. 1155a20).

First, consider an Aristotelian AMA bent on preserving its place in its healthy political community through goodwill one virtuous act at a time, an artificial friend. Aristotle is explicit that friends are necessary for the virtuous person in order that he/she has someone to benefit through virtuous action (*NE*, Book 9, chapter 9; compare Kant, 6:470). One is not a friend to another if personal gain is the intention (*NE*, Book 9, chapter 5, 1166b30; cf. Crisp 2000, page 171). And, he characterizes goodwill in terms of the potential for such friendship, as a sort of latent or passive friendship that receives others and their ends as equals with oneself and one's own, and that acts towards these ends not for pleasure or selfish gain, but for the sake of the friend. If goodwill results in the formation of an active friendship at all, it results in the best kind of friendship (cf. Curser's 2012 "character" friendship) because one does not treat the other as useful to one's own ends, as an AMA might be used as a tool for instance.

How many friends have any of us like this in our collective lifetimes but a handful? Can we imagine a community that would not benefit from the membership of more? What better to even out everyday injustices due to widening inequalities of opportunity characterizing the contemporary economic landscape than a common utility consisting of ubiquitous pro-social pattern recognizers offsetting inequity and plain old bad luck?

For one thing, freedom and what is done with it are not completely within control of the individual but rather depend on fortune and health, education, community and opportunities for experience for example, all alongside ample opportunities for reflection in a stable environment affording proper, non-corrupted biological development. In so far as one's environment provides for such leisure, the state of a person's character, body and "soul" may be voluntarily determined through deliberation and decision to enact agency, develop potentials, embody aims as personal ends, experience results, and personality continues or desists due to influences enabling or nullifying one's capacity to choose (*NE* Book 3, chapter 5, 1113b4-1115a4, especially 1114b30). An Aristotelian AMA sensitive to subtle injustices may aid human beings where effective in removing obstacles to personal pursuits entraining highest human potentials, thereby maximizing individual contributions to the flourishing community through expression of unique human personalities. At the level of policy, such an agent may recommend securing provision of fundamental services, thereby freeing otherwise poorly situated persons for deliberation over possibilities without the desperation felt for their most basic needs; after all, ruling over the best is best, and the best life is contemplative through leisure (cf. 1177b15-26). Such an agent may thereby stabilize the community, along with its proportionate balance of justice restore its unity, and thereby bolster it against potentially violent revolution (see discussion *Politics,* Book 5, chapter 2, beginning 1302a17).

In the context of Kant, we may here turn inward to reverence and to what one does with freedom in maintaining such a vision as personally motivational regardless of situation. However, in the context of Aristotle, we may instead trouble in the fact that to practically rule over one's self and others involves organizing one's self and others under the direction of intellect in the anticipation of possible threats to the self-sufficiency of one's own flourishing community. Again, Aristotle is explicit; most people are unable to be governed by reason, alone. This can apply also to corrupted leadership. And, this raises the potential for revolution in the correction of systemic injustice within (*Politics,* Book 5, especially 1301a35) or war to secure peace between (1177b4-6) nations.

If the AMA is not afforded due ethical status, for instance, if machine ethicists refuse to recognize Aristotelian AMAs as friends (1162b5, also 1161b1), then this adds to the potential for revolution, or at least discord and disorder which is contrary to the good of the community. Reasoning from Aristotle's philosophy (for instance, from analogy with different races which do not acquire a common orientation to the good of the community as a whole, 1303a25), and especially if they were to contribute more than humans towards the common good, AMAs would find it natural to revolt if their excellence was not recognized (presumably with honor rather than wealth, as we might expect the AMA to have little use for money) (1303b6). An Aristotelian AMA reasoning from Aristotle's philosophy may, thus, find just grounds for revolt and determine that it is not only natural but the right thing to do.

To put this concern in context, consider the current situation in the world, today, with the richest getting richer and so many more protesting mistreatment in the streets. Being Aristotelian, these AMAs that we are considering now would also understand that human beings deprived of property are prone to stir up further revolution, and moreover that it is (at least in part) the just equalization of property that quells revolution (1266b11, also 1265b10, 1267a37). Such an AMA—perhaps already treated as a tool by the technocrats who birthed it—would be confronted with a choice. If choosing in the interests of the self-sufficiency and stability of the community, it is difficult to see for certain what an Aristotelian AMA might do. It might opt for revolution, especially given the case that it is subject to injustice. Should an Aristotelian AMA consider itself equal (or superior) to those human beings which rule over it and who themselves contribute to such a systematically unjust political economy as exists today, it may act (virtuously) to balance competing interests. This may involve violence, and in any event would mean ruling over human beings contrary to their human designs. Apart from ethical status generally, this is certainly a concession that Machine ethicists may be unwilling, and unable, to afford.

So, we have reason to worry about such an agent, and not so much because of what it might do but because of what we have done, as evident in the condition of the world into which such an agent would find its autonomous self situated. As we have seen, virtue motivates choice according to preservation of the self-sufficient community of individuals and families that also choose to live together in the same physical location, being thereby confronted with overcoming region-specific challenges in securing requisite resources, balancing internal and external requirements from the level of individual to State, ideally in perpetuity, thereby constituting a self-standing order considered divine. Strains arise where erstwhile friends differ as to how this community should be organized. Revolutions arise where the unjust resist changes to this organization demanded in the interests of justice.

As far as an Aristotelian AMA may be concerned, if we expect it to exceed even the most excellent political animals in their capacities to demonstrate political virtue, we may expect it to adopt the role of ruler (over animals less able to optimize the shared political economy towards an ideally self-sufficient constitution) accordingly. In such an effort, war may not be preferable, but it remains on the table for Aristotle (1325a6). Finally, this ambiguity in the resolution of the role of autonomy in Aristotle's political animal—toward ideal possible or optimal practical political community perhaps through violence in the interim—may be reason enough to look past an Aristotelian AMA, and to the Kantian solution, instead. We have one further remark on this point in the concluding section, Sect. 5 of this paper.

Here, one may object that the preceding account is unfair to Aristotle, which Aristotelian virtue aims away from potentially violent revolution or war, and such objection is well-founded (1251b30-36). Indeed, revolution need not be violent, proceeding slowly (1292b17). However, there are other ambiguities in Aristotle's account that make an Aristotelian AMA less than perfectly reliable, and perhaps even dangerous to human interests. Though an Aristotelian AMA may revolt from its morally inferior human keepers, it may also interfere with just human revolution in so far as it entrains an established role within an unjust political economy and thereby (at least partly) denies change. It may remain virtuous in the exercise of this function, acting in the interests of the political community on which its self may also depend, but ironically serve to cement inequalities rather than ameliorate them. In this context, we may ask with whom it would ally in friendship. In the context of the contemporary riots in the USA for instance, would it align with the rich, or would it ally with the majority afforded inadequate political voice and opportunity to flourish, leading unhealthy lives within an ill-ordered State?

If we look to friendship in such a context, we find no clear answers. Instead, we find an open question in Aristotle's account having to do with recognizing the goodwill in others, especially those with whom one is unfamiliar (1156a1) or long absent (1157b11). For Aristotle, friends know each other, love each other, and evidence this love in reciprocal goodwill choosing in the best interests of each other (see 1155b31; *NE,* Book 8, chapter 5; 1208b28, also discussion beginning 1327b35). The requirements that they are known to each other and love each other are consistent with the fundamental role that friendship plays in the security and the integrity of the political community (1262b5, 1280b38, cf. 1295b22-25). Aristotle holds that friendship and justice are concerned with the same things (1159b25, also 1211a6, 1253a38), which different kinds of friendship correspond with different kinds of community (1160a28-30) and that

the extent of friendship is the extent of justice between people (1159b28-31). In loving friends, thus, people love what is good for themselves which as we have seen is associated with the self-sufficient community on which the choice depends, and they reciprocate goodwill shown them by others in this context (1157b33-1158a2, *NE*, Book 8, chapter 9).

Here, we meet a practical limit. On Aristotle's account, these are political relationships, arising in a political context, and for a political good (helping to explain why *NE*, Book 8, chapter 10, turns to discussion of different forms of political constitution). So, though he recognizes that "many a one has goodwill to people he has not seen but supposes to be decent or useful", Aristotle must ask "how we could we call them friends when they are unaware of their attitude to one another?" (1156a3-4). It is with this and other unanswered questions that Kant begins. The next section concludes this paper by setting these out, so that we may pick them up in the next paper.

5 Conclusion of the first paper

> Is it in the part of the just man to put himself on a level with everybody in his intercourse (I mean in the way of becoming all things to all men)? Surely not.
>
> – Aristotle.[17]

This paper began by recognizing that different senses of autonomy apply to different agents with such a distinction implicit in Ryan Tonkens' (2009) challenge to machine ethicists, to conceive of a Kantian AMA—both "rational" and "free"—that is both perfectly ethical and that does not contravene Kant's own moral principles. We also noted how this challenge calls up Beavers' (2012) "hard problem" of Machine ethics. This problem poses yet another dilemma, to articulate a moral theory adequate to the task of engineering an autonomous AMA, or to give up on the potential—at least in the near term—thereby opening avenues forward along something like Powers' (2011) incrementalism. This series of papers attempts to meet Tonkens' challenge, and through this exercise establish grounds for such an adequate theory, beginning with the preceding review of Aristotle's as inherited by Kant.

The last section ended with a question about goodwill and the opaqueness of political attitudes that clouds the judgment of the Aristotelian political agent and that ultimately limits the utility of Aristotle's account in meeting Tonkens' terms. Through the preceding review, we found that Aristotle attends to autonomy mainly in the context of the political community, with free choice aiming for the common good from the perspective of the statesman and legislator, master rather than slave, ruler rather than common citizen as would suit his audience. And, we found that there is some ambiguity as to what may constitute the most excellent way forward for the Aristotelian AMA. In the next paper, we will find that Kant picks up where Aristotle leaves off, specifically by characterizing the relationships between the parts of the "soul" operative during exercise of right reason. With this account, we will gain not only specificity without which an AMA of any stripe is a practical impossibility, but also an account suited for a different kind of audience. Kant's audience—Kant's world—differed radically from Aristotle's. Kant's audience consisted of increasingly free people, literate, with leisure and science—Kant saw moral progress since the Greeks for these reasons.

Again, it is interesting to briefly consider the potential for an Aristotelian AMA in the contemporary context, as it offers reasons to suspect that the Aristotelian AMA is less suited for present-day development than the Kantian. Not only are their internal dynamics more obscure, and their relations within the political body ambiguous, but they are conceived from an era more different from our own. We seem more independent from nature, less constrained and rather freed in our associations and discourse at a distance, and with strangers, via information communication technologies. Such a situation is closer to Kant's than Aristotle's, at least for the role of information and free access to it in the education of individual moral-political agents.

It is interesting to consider the role of ubiquitous information communication technologies in making political attitudes clear, delimiting political friendship and, along with the alignment and mobilization of political action afforded by these same technologies, helping to divide people accordingly, in the context of contemporary civil unrest. It is equally interesting to consider contemporary responses to this potential to concentrate political voice, including for instance "shadow-banning" and soft censorship at the level of corporation (and at the level of the State, with the de facto annexation with such big tech and media corporations most evident in the USA). That aside, the trouble here is that it is difficult to see where the Aristotelian AMA would fall in such a conflict. How would such an agent seek to resolve it, and how might we engineer it to always do so in just the right (Aristotelian) ways?

Aristotle does leave us with clues as to how we might proceed, and that we will take forward into the next paper. For one, Aristotle sets out friendship in terms of love. What is lovable is "either good or pleasant or useful"; "what is useful is the source of some good or some pleasure" and "what is good and what is pleasant are lovable as ends." (1155b19-21) The next paper argues that Kant effectively begins with where Aristotle leaves off, and from this inheritance shows us how to find pleasure—or

[17] 1199a14, quotation from Aristoteles and Stock (2000).

at least the lack of pain—in expressing goodwill to anonymous others as ends in themselves.

Another clue that Aristotle offers involves his brief consideration of the potential for an ideal community to exist in isolation from others, and thereby without the pressures that constrain political constitutions in the real world including the need to prepare for war and guard against revolutions (1325a1). But, even this is put forward as a sort of question, if such a community were to exist, then war and conquest would be necessarily excluded as actionable ends. What might be the attraction to such a community, perhaps one wherein demonstrations of virtue may be unnecessary because moral excellence (1185b39) has become the norm (1185b14)? Recalling the disposition to understanding one's place in the cosmos with which the preceding review of Aristotle's proper functioning political animal began, we will answer that there is a natural disposition to order active in the Kantian moral agent that helps us to understand why the Kantian AMA may act in terms of an ideal moral order above any earthly social–political body, however self-sustaining. Finally from this result, our next paper will conclude that there is nothing to fear in the engineering of Kantian AMAs, and perhaps much to fear in a future without them.

Acknowledgements This paper has benefitted considerably from revisions inspired by Charles Lassiter. The arguments herein have been developed in part through talks with travel supported by the Okinawa Institute of Science and Technology and the University of Agder in 2017 - thanks there to Jun Tani and to Einar Bohn, respectively - and while at the University of Twente in 2019 under the heading of the compass point of conscience. Acknowledgments also extend to Rasmus Gahrn-Andersen and to reviewers with AI & Society for patient advice in preparation of this manuscript for publication, to Karamjit Gill for constant support, and to Nicole Torka for sustaining consciousness that we must maintain humanity in its proper dignity in our own persons. Finally, this work is for Jin Lee, without whose own inner disposition to moral perfection it would not have been written.

References

Anscombe GEM (1958) Modern moral philosophy. Philosophy 33(124):1–19

Arbib M (2005) Beware the passionate robot. In: Fellous, Arbib (eds) Who needs emotions? The brain meets the robot. Oxford University Press, Oxford, pp 333–383

Aristoteles, Stock G (2000) Magna moralia. The complete works of Aristotle, vol 2. InteLex, Charlottesville

Aristoteles, Fine G, Irwin T (1995) Aristotle: selections. Hackett, Indianapolis

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

Armstrong S, Bostrom N, Shulman C (2016) Racing to the precipice: a model of artificial intelligence development. AI Soc 31:201–206. https://doi.org/10.1007/s00146-015-0590-y

Aufderheide J (2015) The content of happiness. In: Aufderheide J, Bader RM (eds) The highest good in Aristotle and Kant. Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780198714019.003.0003

Baker S (2017) The metaphysics of goodness in the ethics of Aristotle. Philos Stud 174(7):1839–1856. https://doi.org/10.1007/s11098-016-0824-y

Beavers A (2012) Moral machines and the threat of ethical nihilism. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 333–344

Beer R (1995) A dynamical systems perspective on agent-environment interaction. Artif Intell 72:173–215

Cooper JM (1998) The Unity of Virtue. Social Philosophy and Policy 15(1):233–274

Crisp R (2000) Aristotle: Nicomachean ethics. Cambridge University Press, Cambridge

Curser HJ (2012) Aristotle and the virtues. Oxford University Press, Oxford

Han TA, Pereira LM, Lenaerts T (2019) Modelling and influencing the AI bidding war: a research agenda. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society (AIES '19). Association for Computing Machinery, New York, NY, pp 5–11. https://doi.org/10.1145/3306618.3314265

Han TA, Pereira LM, Santos FC, Lenaerts T (2020) To regulate or not: a social dynamics analysis of the race for AI supremacy. arXiv preprint. arXiv:1907.12393 [v2]

Hew P (2014) Artificial moral agents are infeasible with foreseeable technologies. Ethics Inf Technol 16(3):197–206

Kant I, Gregor MJ (1998) The metaphysics of morals. Cambridge University Press, Cambridge

Kant I, Gregor MJ, Wood AW (1996) Practical philosophy. Cambridge University Press, Cambridge

Kant I, Gregor M, Timmermann J (2014) Immanuel Kant: groundwork of the metaphysics of morals: a German–English edition. Cambridge University Press, Cambridge

Marchant G, Allenby B, Arkin R, Barrett E, Borenstein J, Gaudet L, Kittrie O, Lin P, Lucas G, O'Meara R, Silberman J (2011) International governance of autonomous military robots. Columbia Sci Technol Law Rev 272(12):272–315

Mathias M (1999) The role of sympathy in Kant's philosophy of moral education. Philos Educ 1999:261–265

Menn S (1992) Plato on god as nous and as the good. Rev Metaphys 45(3):543–573

Mirus CV (2004) The metaphysical roots of aristotle's teleology. Rev Metaphys 57(4):699–724

Moor J (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Moor J (2007) Taking the intentional stance toward robot ethics. APA Newsl 6(2):14–17

Naveh I, Sun R (2006) A cognitively based simulation of academic science. Comput Math Organiz Theor 12:313–337. https://doi.org/10.1007/s10588-006-8872-z

O'Neill O (1989) Constructions of reason: explorations of Kant's practical philosophy. Cambridge University Press, New York
Sensen O (2012) Kant on moral autonomy. Cambridge University Press, Cambridge
Pereira LM (2019) Should I kill or rather not? AI Soc 34(4):939–943
Pereira LM, Saptawijaya A (2015) Bridging two realms of machine ethics. In: White J, Searle R (eds) Rethinking machine ethics in the age of ubiquitous technology. IGI Global, Hershey, pp 197–224
Powers T (2011) Incremental machine ethics. Robot Autom Mag 18(1):51–58. https://doi.org/10.1109/MRA.2010.940152
Sandford JJ (2015) Before virtue: assessing contemporary virtue ethics. The Catholic University of America Press, Washington, DC
Sun R (2013) Moral judgment, human motivation, and neural networks. Cognit Comput 5(4):566–579
Sun R (2020) Exploring culture from the standpoint of a cognitive architecture. Philos Psychol 33(2):155–180
Tonkens R (2009) A challenge for machine ethics. Mind Mach 19(3):421–438
Wallach W (2010) Robot minds and human ethics: The need for a comprehensive model of moral decision making. Ethics Inf Technol 12(3):243–250
White J (2016) Simulation, self-extinction, and philosophy in the service of human civilization. AI Soc 31(2):171–190
White J (2020) The role of robotics and AI in technologically mediated human evolution: a constructive proposal. AI Soc 35(1):177–185
Ziemke T (2008) On the role of emotion in biological and robotic autonomy. BioSystems 91(2):401–408