Probing Networked Agency: Where is the Locus of Moral Responsibility?

Introduction

As the enterprise of artificial intelligence continues to permeate our sociocultural, political and economic spheres, it becomes increasingly crucial to deliberate over and negotiate its place in our ethical and legal frameworks. The ethical and existential questions posed by the advent of intelligent machinery are not only theoretically insightful, but have direct pragmatic relevance and can pose rather injurious implications if approached maladaptively. In order to mitigate such risks, it is essential that we reflect on our convictions concerning where the onus falls in morally charged cases, so that they can be brought to bear in matters of ethics, law and policy.

Within this essay, I intend to problematize and expand on an analogy entertained by John P. Sullins within his paper "When Is a Robot a Moral Agent?" - one which is drawn between a trained service dog and an intelligent machine, and which concomitantly implores us to consider whether both would assume an equal degree of moral responsibility. While Sullins does not endorse this analogy outright, it is nonetheless a contentious proposition to consider in light of the scholarly debate about the locus of moral responsibility in ethically fraught contexts. I argue that a disanalogy holds between canine and robot in this case: while the canine can be rightfully positioned as a node of moral responsibility within a broader network of interrelated moral agents, the same designation cannot be attributed to the machine. I will begin by contending that while both can be considered "manipulation(s) of nature to human ends", these "manipulations" are vastly different in essence: while dogs possess certain dispositional traits which *precede* human intervention, the behavioural nature of intelligent machinery is engineered and fashioned by humankind such that it is a *direct precipitate* of human intervention. Secondly, I will expound the counterintuitive nature and injurious potential of such an analogy, should it be deployed in practical contexts. Finally, I will sketch a series of hypothetical objections, along with a set of corresponding replies, in order to further dialogue with and underscore the problematic nature of this analogy, as well as to undermine the case for robots as partial loci of moral responsibility.

Networked Moral Agency and the Canine-Machine (Dis)analogy

Within his paper "When is a robot a moral agent?", John P. Sullins aims to establish a conceptual framework of moral responsibility in contexts that implicate multiple potential moral agents; some of these agents being intelligent machines. A number of hypothetical concerns are raised and explored within Sullins' paper, including whether moral rights and responsibilities ought to be administered to agents *by proxy* of intelligent machinery, or, alternatively, whether they ought to be conferred onto the machine itself. In exploring how fellow moral theorists have approached the domain of technoethics, Sullins touches on a number of theoretical moral schemas, including the traditional User, Tool and Victim Model. Within this model, the "technology mediates the moral situation between the actor who uses the technology and the victim" (p. 152). Sullins ultimately denounces this model as overly simplistic, and instead elects to expand our conception of technology in order to challenge our understanding of its passive role within our customary schema of moral responsibility. In his analysis, Sullins argues that humanity has long bred and deployed dogs for anthropocentric ends, and "if we think of technology as a manipulation of nature to human ends, we can comfortably call domesticated dogs a technology" (p. 152).

In order to bolster this argument, Sullins institutes the example of service dogs for the visually impaired. Within such contexts, Sullins notes that the practical and therapeutic merit of the guide dog is often attributed not only to the trainer, but also to the dog itself, with both being morally revered in the process. Here, unlike the attribution of moral agency under the User, Tool and Victim Model, moral agency is thought to be dispersed among a web of interrelations between a host of moral agents, including the breeder, the trainer, the training program developer, the user, and - perhaps more contentiously - the dog itself. Rather than occupying a passive and instrumental role by which human subjects channel their moral agency, the dog, in this case, is thought to take on a more *active* capacity within this network of interrelated agents. Consequently, Sullins considers the guide dog as morally praiseworthy in itself, thus affixing to it a certain degree of moral responsibility.

Sullins further contends that there exists a theoretical parallel between canine and robot in this case, in that these dogs prove quite similar to the automated intelligent machinery that we endeavour to bring about. Insofar as said machinery is as functionally (and affectively) embroiled in this web of interrelations as is the service dog, it should also be dispossessed of its tool-like designation and instead afforded a status of partial moral responsibility. In order to illustrate this point, Sullins points to the advent of autonomous robots, which have an important relationship with moral agency. In view of the fact that autonomous robots possess “a significant degree of autonomous ability to reason and act on those reasons” (p. 154), it can be argued that the machine’s programmer cannot soundly be deemed the sole locus of moral responsibility within morally charged contexts. Rather, this responsibility - at least to a partial degree - ought to be affixed to the machine itself.

Conversely, I argue that while both the service dog and the autonomous robot can be considered “manipulation(s) of nature to human ends”, these “manipulations” are vastly different in essence. Although certain similarities obtain between these subjects, Sullins, in his comparison, negates a fundamental disanalogy which holds between them: while dogs possess certain dispositional traits which *precede* human intervention, the behavioural nature of intelligent machinery is engineered and fashioned by humankind such that it is a *direct precipitate* of human intervention. This disanalogy can be illustrated with the following line of reasoning:

P1: Service dogs are equipped with certain behavioural dispositions which precede human intervention, while machinery is a direct precipitate of human intervention.

The existence of the canine's antecedent behavioural nature not only appeals to our intuition, but is also legible in our ability to distinguish between environmental and hereditary influence on domestic canine behaviour. To elucidate, many would agree that the behavioural tendencies of these animals vary according to morphology rather than human-governed environmental influence. This is evidenced by the significant intra-breed behavioural similarity and inter-breed behavioural variability displayed by different canine breeds: while a fair degree of dispositional consistency is often displayed among members of a common breed, a lesser degree of consistency is observed among members of separate breeds, irrespective of the domestic circumstances to which they are subjected. The discrepancies in environmental influence and perhaps counterintuitive persistence of behavioural similarity among members of common breeds (despite said environmental discrepancies) perhaps illustrate that canines possess a behavioural nature which subsists independently of human influence. The same conclusion, however, cannot be drawn in the case of intelligent machines, whose behavioural nature simply cannot be disentangled from human influence.

P2: Possession of a behavioural nature independent of human imposition is often considered a quality which affords one a morally responsible designation, as it connotes some semblance of agency, free will and conscious volition.

P3: Any being *lacking* a behavioural nature preceding human intervention, or a form of comportment "untainted" by the imposition of humankind, is exempt from moral responsibility (and thereby does *not* constitute the locus of moral responsibility within networks containing multiple subjects).

C1: Service dogs can be ascribed a morally responsible designation and thus stand as one of many nodes of moral responsibility within a broader network of interrelated moral agents. There is not one singular locus of moral responsibility, but many loci; one being the canine.

C2: An intelligent machine cannot be ascribed a morally responsible designation and is thereby exempt from the network of moral responsibility.

Deploying the Canine-Machine Analogy: Injurious Outcomes

In addition to the theoretical tensions inherent in the canine-robot analogy, it is equally imperative to consider its counterintuitive and injurious potential in practical and legislative contexts, which further underscores its problematic nature. With the rise of technological progress comes a paradigm shift not only in how we deploy, but conceptualize intelligent machinery. One central component of this shift may be the designation of these machines as autonomous moral agents, or AMAs. Characterized by a capacity to "reason about the moral and social significance of their behaviour and use their assessments of the effects their behaviour has on sentient beings to make appropriate choices" (Noorman, 2018), this designation will likely only augment in popularity, given machinery's augmenting capacity to operate independently of immediate human control. This is especially likely if the kind of thinking encouraged by the canine-robot analogy were to be adopted by the general public.

Once ethical decision making is inscribed into the robot's operating system, it would not be outlandish for the general populous to ascribe to it a greater sense of moral responsibility and, by extension, dispossess otherwise culpable human agents with command over this robot of their rightful charge of moral responsibility. Consequently, if the canine-machine analogy were to be deployed in legislative contexts, intelligent machines could be exploited to leverage the malicious interests of certain individuals while absolving them from (currently) lawful punishment. These dire outcomes are multiply realizable, and could manifest in contexts ranging from the technological hijacking of a spouse's self-driving car or home service robot in order to execute them and acquire their life insurance benefits, to contexts of robotic warfare, in which quarrelling nation states may, in a time of political unrest, deploy autonomous bomb disposal robots in a fashion that would otherwise be unlawful. However, the general confusion surrounding the ascription of moral responsibility (and the possible tendency to attribute it to the bots under the canine-machine analogy) may present an opportune moment for certain individuals to evade accountability for their nefarious actions.

As a fair deal of moral responsibility could potentially be delegated to the machine, these individuals may feel partially justified in their morally reprehensible practices. This emboldening effect may not only consequently exempt wrongdoers from rightful prosecution or from facing a punishment proportionate with the thrust of their misconduct, but also incentivize further enactments of violence, malice and human rights violations. Thus, if Sullins' analogy were to be deployed in legal contexts, it may passively promote a defiance of our existing moral principles and, by extension, reify a low standard of morality. With the ability to evade culpability, others may opt to neglect their moral convictions and exploit these technologies for selfish gain as well.

Hypothetical Objections and Replies

1. Human Manipulation of Canine Behaviour as an Instantiation of Programming

There exist a host of hypothetical objections to the assertion that the machine is less worthy of a morally responsible designation than the canine. Firstly, one might argue that insofar as service dogs (like intelligent machines) are “programmed” by humans, the difference between them - and, by extension, the machine’s lesser prescription of moral responsibility - no longer hold. Proponents of this stance may emphasize certain modes of human intervention that could arguably render the canine a product of “programming” in its own right. This “programming” can be constituted either in the form of *operant conditioning*, whereby approximations of desired behaviour are reinforced until said behaviour is realized on a consistent basis; or in the form of *artificial selection*, whereby canines exhibiting certain phenotypical or behavioural traits deemed favourable are subjected to a selective breeding process which, under the dominion of human agency, ensures the replication of such traits. Regardless of their differences, both modes of human intervention constitute an authoritative, deliberate and methodical mapping of canine behaviour, which significantly mirrors the mapping of machine behaviour executed through coding or programming. Thus, insofar as both canine and machine can be considered products of “programming” in their respective ways, any significant difference established between the genesis of their behaviour - and, by extension, the maxim to afford them unequal degrees of moral responsibility - collapses.

However, despite the presence of human manipulation in the context of canine training, it is crucial to draw a distinction between the ostensible "programming" of canine behaviour and the programming which undergirds machine behaviour. While modes of intervention like operant conditioning and artificial selection can be positioned as kinds of "programming" in their own right, it must be considered that they require manipulation of a pre-existent genetic or biochemical substrate of behaviour. Here, human agents work with a set of preconceived biological materials; materials with their own essences, propensities and limitations for malleability. Even with the help of the most refined biotechnological strategies, some of these traits are simply too granular in scope for (and ultimately extend beyond the capacity of) human manipulation. Unlike technological programming, the "programming" involved in canine training presupposes an antecedent nature; an essence which is to be taken up, tempered, repressed and "sculpted" into its desired form. Albeit malleable, this nature can never be entirely obscured or annihilated, regardless of the rigour of training to which it is subjected. Machines, on the other hand, are programmed at conception. In the context of technological engineering, it is humankind which fashions these machines' constituent materials and properties. As opposed to the canine, who exhibits some manner of unpredictable and sporadic behaviour, the machine does not have a primal nature which is then fashioned anew. Contrarily, its "nature" is entirely engineered by humankind. As its developmental trajectory and behaviour are grounded entirely in human deliberation, there remains no room for moral freedom or responsibility on the part of the machine.

2. The Argument from Machine Learning

One hypothetical objection to the assertion that while dogs possess certain dispositional traits which *precede* human intervention, the behavioural nature of intelligent machinery is a *direct precipitate* of human intervention is that it may not hold in the context of a robot that develops its behavioural nature through machine learning as opposed to meticulously prescriptive programming. Here, a parallel can be drawn between machine learning and the training undergone by canines, who are thought to claim a degree of moral responsibility in their actions. Machines designed with the machine learning approach can, via experience, effectively re-fashion and optimize their behaviour as needed at least somewhat autonomously, rather than under the complete sovereignty of a human agent. In this sense, the machine is comparable to the canine insofar as their behaviour, too, is subjected to some manner of "training", but is also emergent from some degree of ostensible autonomy. In contexts of machine learning, there is *something* being sculpted; something which is perhaps comparable to a canine's preexistent behavioural tendencies. Designing robots with the machine learning model arguably affords them a capacity to meditate on their actions, the risks posed by these actions, and where these actions may stand in our pre-established moral framework. Consequently, machine learning can be conceptualized as a circumstance in which some magnitude of moral responsibility is delegated from the human agent to the robot.

However, although their learning process seemingly occurs independently of human agency, it is crucial to note that the robot's "learning style" remains inscribed in its human-fashioned behavioural blueprint, unlike that of the canine. This is illustrated in Johnson and Verdicchio's text "Why robots should not be treated like animals" (2018), in which it is argued that "differently from what happens in genetics (which poses limitation to canine "behaviour-sculpting"), humans do have a complete knowledge of the workings of the electronic circuitry of which a robot's hardware is comprised, and the instructions that constitute the robot's software have been written by a team of human coders. Even the most sophisticated artefacts that are able to learn and perfect new tasks, thanks to the latest machine learning techniques, depend heavily on human designers for their initial set-up, and human trainers for their learning process." (p. 297). Even in the process of bots learning and making recursive improvements to their own behaviour, this behaviour transpires in accordance with, and is ultimately grounded in a set of circumscribed design conditions determined by the programmer. These robots' freedom of action can thereby only extend to the contours of these conditions; conditions which are ultimately governed by human agents. Thus, insofar as robot behaviour will always have its genesis in human innovation and choice, the onus for said behaviour should always fall on the human innovator and/or user. By extension, autonomous robots should be classified as morally neutral products of human innovation rather than as volitional agents with an instinctive agenda. As humankind has designed and configured all of its constituent parts, it can be argued that machine behaviour is far more amenable to demystification and control than is canine behaviour, and that the robot cannot reasonably be afforded the same space for behavioural autonomy or moral responsibility as the canine.

3. The Argument from Natural Selection

Another argument to the assertion that while dogs possess certain dispositional traits which *precede* human intervention, the behavioural nature of intelligent machinery is a *direct precipitate* of human intervention is that the canine's "antecedent, preexistent nature" (which affords it partial moral responsibility) can be considered a product of design in its own right. This is because it is fashioned and governed by nature - specifically, via the evolutionary mechanism of natural selection - in the same sense that technological behaviour is fashioned and governed by humankind. This would appear to undermine the canine's volitional potential, rendering its status as a moral agent no more tenable than that of a machine. As the respective "natures" of canine and robot are ultimately "designed" (by the biological mechanism of natural selection and by human programming, respectively), the distinction drawn between these natures is revealed to be devoid of explanatory value, and fails to further the case for dogs as being more viable morally responsible agents than robots.

In response to this criticism, it can be contended that even if the designation of the canine's nature as a product of "biological design" carried out under the dominion of nature proves to be true, there are no means of ascertaining whether this process is grounded in any manner of *forethought* or *intent*, as is the technological design of machine carried out by human agents. Given that forethought and intent are intersubjectively valid criteria for moral responsibility, so long as the question of an external and supreme force *intentionally* directing the natural, antecedent behaviours of the canine is up for debate, we cannot soundly ascribe responsibility to anything/one other than the canine itself. It is not the property of being "designed" which exempts one from a morally responsible designation, but rather the property of being designed by a *demonstrably volitional entity* with specific and demonstrable *intent.*

Conclusion

In conclusion, the analogy entertained by John P. Sullins within his text "When is a robot a moral agent?" is doubly problematic. Firstly, it fails to acknowledge a fundamental disanalogy which holds between canine and robot: while the former can be rightfully positioned as a node of moral responsibility within the broader network of interrelated moral agents, the same designation cannot be attributed to the latter. Although both can be considered "manipulation(s) of nature to human ends", these "manipulations" are vastly different in essence: while dogs possess certain dispositional traits which *precede* human intervention, the behavioural nature of intelligent machinery is engineered and fashioned by humankind such that it is a *direct precipitate* of human intervention. Secondly, Sullins' analogy has rather injurious potential when considered in light of hypothetical practical contexts; namely those in which human agents might evade culpability by appealing to the supposed moral agency of bots deployed as vehicles for their own destructive agendas. According to Noorman (2018), "Progressively autonomous technologies already in development, such as military robots, driverless cars or trains and service robots in the home and for healthcare, will be involved in moral situations that directly affect the safety and well-being of humans." Several questions and concerns emerge from this prospective state of affairs: will it ever be possible for the agency of robotic entities to transcend human influence, and if so, would this agency be amenable to quantification? How might this affect where we perceive the onus to fall in morally charged contexts? Who (or what) should be held liable, and to what degree? As autonomous technologies advance in complexity and sophistication, so too must our ethics and legislation. In light of the currently unprecedented rate of technological progress, it is crucial to contemplate our ascriptions of moral responsibility, along with their legal and ethical stakes.

Works Cited

Johnson, D. G. & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology* 20 (4):291-301.

Noorman, M. (2018). Computing and Moral Responsibility. *The Stanford Encyclopedia of Philosophy.* Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics 6* (12):23-30.