Fabio Tollon

Moral Agents or Mindless Machines?
A Critical Appraisal of Agency in Artificial Systems

Abstract

In this paper I provide an exposition and critique of Johnson and Noorman's (2014) three conceptualizations of the agential roles artificial systems can play. I argue that two of these conceptions are unproblematic: that of causally efficacious agency and "acting for" or surrogate agency. Their third conception, that of "autonomous agency," however, is one I have reservations about. The authors point out that there are two ways in which the term "autonomy" can be used: there is, firstly, the engineering sense of the term, which simply refers to the ability of some system to act independently of substantive human control. Secondly, there is the moral sense of the term, which has traditionally grounded many notions of human moral responsibility. I argue that the continued usage of "autonomy" in discussions of artificial agency complicates matters unnecessarily. This occurs in two ways: firstly, the condition of autonomy, even in its engineering sense, fails to accurately describe the way "autonomous" systems are developed in practice. Secondly, the continued usage of autonomy in the moral sense introduces unnecessary metaphysical baggage form the free will debate into discussions about moral agency. In order to understand the debate surrounding autonomy, we would therefore first need to settle many seemingly intractable metaphysical questions regarding the existence of free will in human beings.

Keywords: moral agency, autonomy, artificial agents, moral responsibility, free will

I. INTRODUCTION

Instead of asking the question of whether an entity is deserving of moral concern, moral agency grapples with the question of whether an entity is capable of moral *action*. An agent is simply a being with the capacity to *act* (Schlosser 2015). A moral action would therefore be a type of action for which evaluation using moral criteria would make sense. Inevitably, this type of discussion leads

to further questions concerning responsibility, as it is traditionally supposed that a moral action is one that an entity can be morally responsible for by being accorded praise or blame for the action in question. This type of moral responsibility has historically been reserved for certain biological entities (generally, adult humans). However, the emergence of increasingly complex and autonomous artificial systems might call into question the assumption that human beings can consistently occupy this type of elevated ontological position while machines cannot. The key issue that arises in such discussions is one of *attributability*, and, more specifically, whether we can attribute the capacity for *moral agency* to an artificial agent. The ability to make such an ascription could lead to the resolution of potential "responsibility gaps" (Champagne–Tonkens 2013; Müller 2014; Gunkel 2017; Nyholm 2017): cases in which warranted moral attributions are currently indeterminate.[1] As machines become increasingly autonomous, there could come a point at which it is no longer possible to discern whether or not any human error could in fact have been causally efficacious in bringing about a certain moral outcome (Grodzinsky–Miller–Wolf 2008. 121).

Of course, it is the *capacity for moral agency* that makes someone eligible for moral praise or blame, and thus for any ascription of moral responsibility (Talbert 2019). Deborah Johnson is one author who has made a substantial and important contribution to discussions surrounding the moral roles machines may come to play in human society. Johnson claims that we should be weary of broadening the set of entities known as moral agents, such that they include machines. Her instrumentalist view of technology holds that technological artefacts are always embedded in certain contexts, and that the meaning of this context is determined by the values of human society. Machines are merely the executors of certain functions, with human beings setting the targets of these functions. She therefore maintains that artificial systems can only ever be moral *entities*, but never *moral agents* (Johnson 2006).

It is with these considerations in mind that I will investigate how the conceptual framework provided by Johnson and her various coauthors can help us better understand the potentially morally-laden roles that, increasingly, autonomous machines can come to fulfil in human society now and in the future. To do this, I provide an exposition of three types of agency that might prima facie be accorded to machines, posited by Johnson and Noorman (2014). Two of these types of agency are seemingly uncontroversial, as they deal with artefacts that operate in functionally equivalent ways when compared to human actions. The third conception, however, is much contested, as it deals with the autonomy of

[1] Conversely, "retribution gaps" may also arise. These are cases where there we have strong evidence that a machine was responsible (at least causally) for producing some moral harm. In such scenarios, people may feel the strong urge to punish somebody for the moral harm, but there may be no appropriate (human) target for this punishment (Nyholm 2017).

the potential agent in question. It is also this sense of autonomy that grounds various notions of *moral* responsibility, and so, in order for agents to be moral agents, they must, supposedly, meet this requirement. Johnson argues that the moral sense of autonomy should be reserved for human beings, while the engineering sense can successfully apply to machines. I will claim that the concept of autonomy cannot refer at the level of the *design* of artificial systems (at least for now) but may plausibly refer at the level of our *descriptions* of such systems. Moreover, I will show how Johnson's specific sense of "moral autonomy" carries unnecessary metaphysical baggage.

II. TYPES OF AGENCY

The metaphysics of agency is concerned with the relationship between *actions* and *events*. The most widely accepted metaphysical view of agency is event-causal, whereby it is claimed that agency should be explained in terms of agent-involving states and events (Schlosser 2015). In other words, agency should be understood in terms of *causation*, and, more specifically, in terms of the causal role the agent plays in the production of a certain event. Agents, therefore, are entities capable of having a certain effect on the world, where this effect usually corresponds to certain goals (in the form of desires, beliefs, intentions etc.) that the agent has.

1. Causally efficacious agency

In the context of potential artificial agency, perhaps the most comprehensible conception of "agency", as put forward by Johnson and Noorman (2014), is that of a causally efficacious entity. This conception of agency simply refers to the ability of some entities – specifically technological artefacts – to have a causal influence on various states of affairs, as extensions of the agency of the humans that produce and use them (ibid. 148). This includes artefacts that may be separated from humans in both time and space (for example, attitude control[2] in a spacecraft in orbit around the earth) as well as artefacts that are deployed directly by a human being. A fair question to raise at this juncture is whether it in fact makes sense to consider these types of artefacts agents at all. One option is to conceptualize them as *tools* instead. The reason for preferring the terminology of "agent" as opposed to "tool" is that these artefacts have human intentions programmed/encoded *into* them (Johnson 2006. 202). This is in contrast to a tool,

[2] Attitude control is the controlling of the orientation of an object with reference to an inertial frame, or another entity (e.g. a nearby object, the celestial sphere, etc.).

such as a hammer, which may be used by someone to perform a specific task but does not have the specifications of this task as part of its very make-up. It cannot in any way perform or represent the task independently of human manipulation. The key distinction then between a tool and a technological artefact, according to Johnson, is that the latter has a form of intentionality as a key feature of its make-up, while the former does not (ibid. 201).[3] In this sense, referring to the intentionality of technology would denote the fact that technological artefacts are designed in certain ways to achieve certain outcomes. Consider the simple example of a search engine: keys are pressed in a specific order in an appropriate box and then a button is pressed. The search engine then goes through a set of processes that have been programmed into it by a human being. The "reasons" for the program doing what it does are therefore necessarily tethered to the intentions of the human being that created it.

It makes sense to think of such artefacts as possessing "agency" to the extent that the ubiquity and specific design of these types of artefacts make a difference to the effective outcomes available to us. For example, they make possible novel means with which to achieve our ends by increasing the amount of potential action schemes at our disposal (Illies–Meijers 2009. 422). These artefacts can therefore be thought of as enlarging the possible range of actions available to a particular agent in a given situation. Yet, while it is clear that artefacts can thus have causal efficacy in the sense that they may *contribute* to the creation of certain novel states of affairs, this causal contribution is only efficacious in *conjunction* with the actions of human beings (Johnson–Noorman 2014. 149). The reason we can think of these causally efficacious artefacts as agents is the fact that they make substantial causal contributions to certain outcomes. In this way the causal efficaciousness of an entity leads, in the form of a non-trivial action performed by that entity, to a specific event.

As suggested earlier, we can legitimately think of these artefacts as agents, due to the fact that their manufacturers have certain intentions (aims) when designing and creating them, and so these systems have significance in relation to humans (Johnson–Noorman 2014. 149). The type of agency that we can extend to artefacts under this conception would thus not be one that involves any meaningful sense of responsibility on the part of the artefact, and, by extension, would not entail a distinctly *moral* type of agency. While Johnson and Noorman concede that artefacts can be causally efficacious in the production of various states of affairs, their (the artefacts') contribution in this regard is *always* in combination with that of human beings (ibid. 149). On this conception of agency,

[3] The intentionality of the program should be understood in functional terms, according to Johnson (2006. 202). What this means is that the functionality of these systems has been intentionally created by human designers, and so is necessarily tethered to and wholly determined by human intentions. Human intentions, in this sense, provide the "reasons" why the technological artefact acts in a particular way.

therefore, we can only consider entities that act in "causal conjunction" with human beings.

The next conception of "agency" that I will unpack can be employed for machines that perform tasks on behalf of, but independently from, human operators and so can be seen as a special case of causally efficacious agency.

2. "Acting for" agency

This conception of agency focuses on artefacts that act on behalf of human operators in a type of "surrogate" role (Johnson–Powers 2008; Johnson–Noorman 2014). In an analogous way, when it comes to human beings, surrogate agency occurs when one person acts on behalf of another. In these cases, the surrogate agent is meant to represent a client, and therefore is constrained by certain rules and has certain responsibilities imposed upon them.[4] This type of agency involves a type of representation: the surrogate agent is meant to use his or her expertise to perform tasks and provide assistance to and act as a representative of the client, but does not act out of his or her own accord in that capacity (Johnson–Noorman 2014. 149). When it comes to artificial systems, this "acting for" type of agency occurs in those artefacts that replace or act on behalf of humans in certain domains. Take the example of a stockbroker: in the past, in order to have a trade executed, one would have to phone a stockbroker and request the purchase/sale of a specific share. The stockbroker, acting on your behalf, would then find a willing buyer/seller in the market and execute the trade. The reality today is much different: individuals can now create accounts on trading platforms and buy shares online without the need of a stockbroker. Furthermore, the exchanges on which these trades are made are also run by computers: inputting a "sell order" places your request in an order book, but this order book is not a literal one, as it might have been in the past, and so there is no need to leave the comfort of your home to perform these tasks. Current online order books are fluid, competitive spaces in which high frequency trading occurs, without the need for humans to keep record, as this job is taken care of by the computer powering the system.[5] Technical details aside, what the aforementioned example brings to light is how tasks that were once the exclusive domain of human beings are now performed by artificial systems without too much "hands-on" human involve-

[4] For example, lawyers in certain legal systems are not allowed to represent clients whose interests may conflict with that of another client.

[5] Another interesting development in automated trading has been the explosion of this technology as it applies to cryptocurrency markets. These markets have been heavily impacted by the emergence of "trading bots" which replace the individual as the executor of a buy/sell instruction. The human operator simply inputs certain key parameters and the bot does the rest.

ment. The function of the tasks performed by these systems, however, is still the same: the purpose of an automated trade is still the same as a trade executed by a human, as in both cases the end being pursued is the purchase/sale of some share at the behest of a given client. What has changed, however, is the means by which that specific end is obtained – the artefact acts within given parameters but does not have each action specifically stipulated by a human operator. Some authors (Johnson 2006; Johnson–Powers 2008; Johnson–Noorman 2014) go on to claim that because of this, these technological artefacts have a greater degree of intentionality than causally efficacious agents do. The causally efficacious agent is simply one that had an influence on outcomes in conjunction with human beings. The "acting for" agent, on Johnson and Noorman's construal, should be understood in terms of an analogy: it can be useful to think of artificial systems *as if* they acted on our behalf (in an analogous way to how a lawyer represents their client), but the decisions they make are not the same as the ones made by human beings. The range of actions available to them is still a direct function of the intentions of their programmers/designers, and is in this sense "determined", whereas human action, according to Johnson and Noorman at least, is not. These agents differ from causally efficacious agents in that they have a greater degree of independence from direct human intervention, and thus have human intentionality modelled into their potential range of actions to a greater degree than the causally efficacious agent does.

Johnson claims that when we evaluate the behaviour of computer systems "there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user" (Johnson 2006. 202; Johnson–Powers 2005).[6] Nevertheless, these artefacts, in order to function as desired, are fundamentally anchored to their human designers and users (Johnson 2006. 202). This is true of systems whose proximate behaviour is independent of human operators, as even in such cases, the functioning of the system is determined by its design and use, and both of these aspects involve human agents. These human agents have internal mental states such as beliefs, desires, etc., and, according to Johnson (ibid. 198), it is here that we locate "original" intentionality (and hence, according to Johnson, the responsibility for any of the system's actions)

If the tasks that are delegated to these kinds of artificial agents have moral consequences, this would provide another way in which to conceptualize the role such artefacts could play in our moral lives (Johnson–Noorman 2014. 155). Consider, for example, automatic emergency braking (AEB) technology, which automatically applies the brakes when it detects an object near the front of the vehicle. This simple system has been enormously successful, and research indicates that it could lead to reductions in "pedestrian crashes, right turn crashes,

[6] Johnson and Powers (2005. 100) refer to this as "Technological Moral Action" (TMA).

head on crashes, rear end crashes and hit fixed object crashes" (Doecke et al. 2012). We can usefully think of AEB as assisting us in being better and safer drivers, leading to decreased road fatalities and injuries. These artificial systems, of which AEB is an example, can therefore be seen as performing delegated tasks which can have moral significance. We can therefore meaningfully think of them as being morally relevant *entities*. However, according to Johnson (2006), because of the type of intentionality these entities have, they cannot be considered to be moral *agents*. Johnson claims that the intentionality that we can accord to technological artefacts is only a product of the intentionality of a designer and a user, and so this intentionality is moot without some human input (ibid. 201). When designers engage in the process of producing an artefact, they create them to act in a specific way, and these artefacts remain determined to behave in this way. While human users can introduce novel inputs, the conjunction of designer- and user-intentionality wholly determines the type of intentionality exhibited by these types of computer systems (ibid. 201). Therefore, while it is reasonable to assess the significance of the delegated tasks performed by these artefacts as potentially giving rise to moral consequences, it would be a category mistake "to claim that humans and artefacts are interchangeable components in moral action" in such instances (Johnson–Noorman 2014. 153). For example, consider the type of moral appraisal we might accord to a traffic light versus a traffic officer directing traffic: while these two entities are, in a functional sense, performing the same task, they are not morally the same (Johnson–Miller 2008. 129; Johnson–Noorman 2014. 153).

In order to press this point further, Johnson and Miller draw a distinction between "scientific" and "operational" models and how we evaluate each one respectively (2008. 129). According to the authors, scientific models are tested against the real world, and, in this way, these types of models are constrained by the natural world (ibid. 129). For example, we can be sure we have a good model of a physical system when our model of this system accurately represents what actually occurs in the natural world. Operational models, on the other hand, have no such constraints (besides, of course, their physical/programmed constraints). These models are aimed at achieving maximum utility: they are designed to realise specific outcomes without the need to model or represent what is actually going on in the natural world (ibid. 129). For example, a trading bot (as discussed above) need not in any way model human thinking before executing a trade. All that is important for such a bot, for example, is that it generate the maximum amount of profit given certain constraints. Moreover, the efficacy of such systems is often exactly that they exceed the utility provided by human decision making, usually in cases where complex mathematical relationships between numerous variables need to be calculated. In light of this, Johnson and Miller argue that because only the *function* of the tasks is the same (when comparing human action to operational models), we should not think of such systems as

moral agents, as this would reduce morality to functionality, an idea which they are directly opposed to (ibid. 129). For now, all that should be noted is that artefacts can be agents that, when acting on behalf of human beings, participate in acts that have moral consequences. This, however, does not necessarily mean they are morally responsible for the actions they participate in bringing about: once again, in the current literature, this responsibility is reserved for human beings. In order to be morally responsible, an agent must also have autonomy.

3. Autonomous agency

The third and final conceptualization of agency to be dealt with is that of autonomous agency. On the face of it, there are two ways in which we might come to understand the "autonomous" aspect of this account. Firstly, there is the type of autonomy that we usually ascribe to human beings. This type of autonomy has a distinctly moral dimension and, according to Johnson and Noorman (2014. 151), it is due to our autonomy in this sense that we have the capacity for moral agency. "True" autonomy is often used in discussions of moral agency as the key ingredient which supports idea that only human beings qualify as moral agents, as we are the only entities with the capacity for this kind of autonomy (see Johnson 2006, 2015; Johnson–Miller 2008; Johnson–Powers 2008; Johnson–Noorman 2014). Hence, it is due to the fact that individual human beings act for reasons that they can claim "authorship" for, that they can be said to be truly autonomous and this is what allows us to hold one another morally responsible for our actions (also see Wegner 2002. 2). According to Johnson and Noorman (2014. 151) if a being does not have the capacity to choose *freely* how to act, then it makes no sense to have a set of rules specifying how such an entity *ought* to behave. In other words, the type of autonomy requisite for moral agency here can be stated as the capacity to *choose freely how one acts* (ibid. 151).

However, there is a second understanding of "autonomous agency" that has to do with how we might define it in a non-moral, engineering sense. This sense of autonomy simply refers to artefacts that are capable of operating independently of human control (Johnson–Noorman 2014. 151). Computer scientists commonly refer to "autonomous" programs in order to highlight their ability to carry out tasks on behalf of humans and, furthermore, to do so in a highly independent manner (Alonso 2014). A simplistic example of such a system might be a machine-learning algorithm which is better equipped to operate in novel environments than a simple, pre-programmed algorithm. Nevertheless, this capacity for operational or functional independence is, according to Johnson and Noorman (2014. 152), not sufficient to ground a coherent account of *moral* agency, since, as they argue, such agents do not freely choose how to act in any meaningful sense. So, while the authors do not suggest we eliminate the stand-

ard convention of speaking about "autonomous" machines, they insist on carefully articulating which sense of autonomy is being used. "True" autonomy, on their view, should be reserved for human beings. We should be sensitive to the specific sense of autonomy we mobilise, as confusing the two senses specified here can lead to misunderstandings that may have moral consequences (ibid. 152).

To see how this might play out, it will be helpful to consider how the conception of "truly" autonomous agency not only grounds morality as such, but also confers a particular kind of moral *status* on its holder (ibid. 155). As stated above, this conception of agency has historically served the purpose of distinguishing humans from other entities. As noted above, the traditional means by which this has been achieved is by postulating that human beings exercise a distinct type of *freedom* in their decision making, which is what grounds a coherent sense of moral responsibility. Freedom in this sense is about having meaningful control over one's actions, a type of control which makes a decision or action *up to the agent* and *not* other external circumstances. It is possible for agents of this kind to have done otherwise – they deliberately and freely choose their actions. Moreover, the sense of freedom described above has a sense of autonomy embedded into its definition: if this free decision is not the product of the specific agent in question, and is rather due to external pressures, then we cannot meaningfully consider the action to be free, and hence we would be hard-pressed to hold the agent in question morally responsible for such an act. An example of such a decision would be if an agent was coerced into performing some action (perhaps by physical force or by psychological manipulation), in which case we would not consider the act to have been performed "freely".

These apparent differences in capacity for autonomous action also influence the types of rights we can coherently accord to various entities. On the basis of being autonomous moral agents, humans are accorded several clusters of positive and negative rights, and differences in the type of moral standing we possess can alter the kinds of rights we are extended (ibid. 155). For example, in democratic states there is a minimum legal voting age. One justification for this type of law is the claim that one should only be allowed to vote when one reaches an age of political maturity: an age at which one can exercise the necessary capacities to *consciously* make a *well-informed* vote. In this instance, one's capacity to make informed – and hence, ostensibly *free* – political decisions, captured in a minimum voting age, comes to inform the type of rights one is conferred (i.e. the right to vote). It is against this background that it is argued that we should be careful to distinguish between the two conceptions of autonomous agency identified here and realise that artefacts should not be understood as having the morally relevant kind of autonomy, as we cannot reasonably consider them to be choosing *freely* how they act. *Their* actions are always tethered to the intentions of their designers and end users.

III. PROBLEMATISING AUTONOMY

To reiterate, Johnson argues that we should be cognisant of the distinction between "autonomy" as it is used in the engineering sense, and "autonomy" as it is used when applied to human beings, especially in the context of moral theorising. The engineering sense refers to how an entity may be able to operate outside human control; the moral sense refers to a "special" capacity that human beings have, elevating us above the natural world and making us morally responsible for our actions. In what follows, I will raise two issues with the continued usage of "autonomy" in discussions surrounding AI. The first issue is more general and applies to the engineering sense, while the second issue is directed at Johnson's specific usage of autonomy in the moral sense. The first issue relates to the *design* of AI systems, while the second relates to the *description* of such systems.

1. Losing the definitional baggage

By "autonomous" what is usually meant is the ability of an entity to change states without being directly caused to do so by some external influence.[7] This is a very weak sense of the term (in contrast with the way it is traditionally understood in moral and/or political philosophy) but the basic idea can be grasped with this definition.[8] It captures the major distinction between how the term is used in the design of AI systems and how it refers when applied to human beings: the engineering sense and the so-called "moral" sense. In AI research, one of the main goals of creating machine intelligence is to create systems that can act autonomously in the engineering sense: reasoning, thinking and acting *on their own, without human intervention* (Alterman 2000. 15; Van de Voort–Pieters–Consoli 2015. 45). This is a design specification that has almost reached the level of ideology in AI research and development (Etzioni–Etzioni 2016). When we use this "weak" sense of autonomy, we are usually referring to how a specific AI system has been designed. More specifically, we aim to pick out a system that is able to act independently of human control.

However, as argued by Alterman (2000. 19), identifying machine autonomy is already problematic, as the distinction between the non-autonomous "getting ready" stage and the autonomous "running" stage in the design of a spe-

[7] Changing states simply refers to an entity's ability to update its internal model of the world by considering new information from its environment. This can be as simple as a thermostat keeping the temperature at a set level despite the temperature dropping in the environment (Floridi and Sanders, 2004).

[8] For example, see Christman (2018) for an exposition on how autonomy refers in the moral and political arenas.

cific AI system is a spurious one at best. In the first "getting ready" stage, a system is prepared for deployment in some task environment. In the second stage, the system "runs" according to its design (ibid. 19). Traditionally, it was supposed that these two stages are what separate the "autonomous" from the "non-autonomous" states of the machine. However, consider a case where a system has completed the "getting ready" stage and is ready to "run", and suppose that while entering its "running" state in its given task environment, the system encounters an error. In such a situation, it would be necessary to take the system back into the "getting ready" stage in the hope of fixing the bug. In this way, there is a cycling between the "getting ready" and "running" stages, which entails cycling between stages of "autonomous" and "non-autonomous" learning (ibid. 19). This means that the system's "intelligence" is a function of both stages, and so it becomes unclear where we should be drawing the line between what counts as autonomous or non-autonomous in terms of the states of the machine. According to Alterman, "if the system is intelligent, credit largely goes to how it was developed which is a joint person–machine practice" (ibid. 20). In other words, if the system is considered intelligent, this is already largely a carbon-silicon collaborative effort. Instead of asking whether the system is autonomous or not then, we should perhaps instead inquire as to how its behaviour might be independent from its human designers. What this entails, for my purposes, is that when talking about the *design* of AI systems we should not talk about "autonomous" AI. If autonomy means "independence from human control" then this concept cannot refer at the level of design, as at this level of description, human beings are still very much involved in moving the system from the "getting ready" stage to the "running" stage. The implications of this for Johnson's argument, therefore, are that we need not worry about any confusion regarding the autonomy of AI systems, as the engineering sense of the term fails to refer successfully.[9]

2. Losing the metaphysical baggage

Johnson claims that there is something "mysterious" and unique about human behaviour, and that this mysterious, non-deterministic aspect of human decision-making makes us "free", and therefore morally responsible for our actions (Johnson 2006. 200). Details of the philosophical debate surrounding free will is not something I would wish for anyone to have to explore in full, but my senti-

[9] This is not to deny that in the future we may come to encompass machines that are autonomous in this sense. This would entail that they are capable of setting their own goals and updating their own programming. My intuition is that this outcome is inevitable, but substantiation of this claim is beyond the scope of this paper.

ment towards this debate is no substitute for an argument. The real issue with gesturing towards human freedom as a way of grounding our moral autonomy is that one then brings metaphysically contested claims from the free will literature into a debate about moral agency. Johnson's claim rests on the fact that she presupposes some form of incompatibilism[10], more specifically libertarianism[11] (about free will, not politics). This is a controversial position to hold and is in no way the generally accepted view in philosophical debates on free will (O'Connor–Franklin 2019). There are philosophers who have spent considerable time arguing against such incompatibilist positions (see Dennett 1984, 2003; Pereboom 2003, 2014; cf. Kane 1996). In order to understand the debate surrounding autonomy, we would therefore first need to settle many (seemingly intractable) metaphysical questions regarding the existence of free will in human beings. In this way, her argument that the "freedom" of human decision making is what grounds the special type of autonomy that we apparently have generates far more problems than solutions.

My claim, therefore, is that this sense of autonomy, as Johnson uses it in her *description* of AI systems, invites confusion. For example, the most common usage of the term "autonomous" in discussions on machine ethics usually revolves around military applications (see Sparrow 2007; Müller 2014). A key issue here, however, can be noted in the metaphysical baggage that comes with the ascription of autonomy to a system. To see how this may play out in actual philosophical discourse, consider the following remark by Sparrow, where he claims that "autonomy and moral responsibility go hand in hand" (2007. 65).[12] On this analysis, any system that is deemed autonomous (for example, a military drone), and were to cause some moral harm, would be morally responsible for this harm (ibid.). This would miss key steps in the analysis, as in such a situation, we skip from autonomy to responsibility without, for example, asking whether the entity is also adaptable (Floridi–Sanders 2004).

Returning to the military drone above: imagine it is sent to execute a strike on a certain pre-determined location. This location is programmed (by a human) into the drone before it takes off, but from the moment of take-off, the drone acts autonomously in executing the strike. Let us assume that the strike is unsuccessful, as instead of terrorists, civilians were at the strike location. In this case, while the drone is autonomous, we would not hold the drone morally

[10] This view claims that the truth of determinism is incompatible with freely willed human action.

[11] Libertarians claim that determinism rules out free will but make the further point that our world is in fact indeterministic. It is in these indeterminacies that human decision-making occurs, with the implication that these decisions are free, as they are not necessarily bound to any antecedent causal events and laws that would make them perfectly predictable.

[12] Note that Sparrow (2007) does explicitly state that he remains agnostic on questions of full machine autonomy.

responsible for this outcome, as the moral harm was due to human error. In this weak sense, the criterion of autonomy would provide an implausible account of agency more generally, as it would never allow for minimally "autonomous" machines that are not morally responsible for their actions. The aforementioned case is clearly an oversimplification of the issue, but what the example brings to light is that our ascriptions of autonomy need not be synonymous with those of moral responsibility. Therefore, it should be clear that autonomy does not also necessitate moral responsibility on the part of the agent. This leaves room for autonomous systems that are not morally responsible for their actions.

I therefore suggest that we keep the concepts of autonomy and moral responsibility distinct. Johnson unnecessarily conflates the two (in the case of humans) in order to show how machines can never be "fully autonomous". This attitude however misses key nuances in the debate surrounding machine autonomy and glosses over the fact that it is possible for some systems to be semi-autonomous (such as the one in the drone strike example). Her specific sense of autonomy (by tethering it to the kind of autonomy exhibited by human beings with free will) also introduces unnecessary metaphysical baggage into the discussion. There are far more naturalistically plausible accounts of autonomy which do not involve such metaphysical speculation. Examples of this could be that autonomy is the ability to act in accordance with one's aims, the ability to govern oneself, the ability to act free of coercion or manipulation, etc. (see Christman 2018 for discussion). It is beyond the scope of this paper, however, to provide a positive account of autonomy. Rather, my purpose has been to critically evaluate the specific conception of the term put forward by Johnson. I leave the crucial work of providing such a positive account to other philosophers.

IV. CONCLUSION

I began by introducing the concept of agency, and, more specifically, that of moral agency. I then provided an exposition of three distinct types of agency that we might reasonably accord to artificial systems. While I argued that the three conceptualizations of agency introduced capture many of the ways in which we can meaningfully consider the roles that artificial artefacts play, I expressed reservations regarding the third one of these, that of the "autonomy" condition in our philosophizing about moral agency. I claimed that the continued usage of such a metaphysically loaded term complicates our ability to get a good handle on our concepts and obscures the ways in which we can coherently think through nuanced accounts of the moral role(s) that machines may come to play in our lives.

REFERENCES

Alonso, Eduardo 2014. Actions and Agents. In Keith Frankish – William M. Ramsey (eds.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge, Cambridge University Press. 232–246.

Alterman, Richard 2000. Rethinking Autonomy. *Minds and Machines* 10(1). 15–30. https://doi: 10.1023/A:1008351215377.

Champagne, Marc – Ryan Tonkens 2013. Bridging the Responsibility Gap. *Philosophy and Technology*. 28(1). 125–137.

Christman, John 2018. Autonomy in Moral and Political Philosophy. *The Stanford Encyclopaedia of Philosophy*. Edited by E. N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford, Clarendon Press.

Dennett, Daniel C. 2003. *Freedom Evolves*. New York/NY, Viking Press.

Doecke, Samuel D. et al. 2012. The Potential of Autonomous Emergency Braking Systems to Mitigate Passenger Vehicle Crashes. *Australasian Road Safety Research, Policing and Education Conference*. Wellington, New Zealand.

Etzioni, Amitai – Oren Etzioni 2016. AI Assisted Ethics. *Ethics and Information Technology*. 18(2). https://doi: 10.1007/s10676-016-9400-6.

Floridi, Luciano – Jeff Sanders 2004. On the Morality of Artificial Agents. *Minds and Machines*. 14. 349–379. https://doi:10.1023/B:MIND.0000035461.63578.

Grodzinsky, Frances S. – Keith W. Miller – Marty J. Wolf 2008. The Ethics of Designing Artificial Agents. *Ethics and Information Technology*. 10(2–3). 115–121. https://doi: 10.1007/s10676-008-9163-9.

Gunkel, David J. 2017. Mind the Gap: Responsible Robotics and the Problem of Responsibility. *Ethics and Information Technology* https://doi: 10.1007/s10676-017-9428-2.

Illies, Christian – Anathonie Meijers 2009. Artefacts Without Agency. *The Monist*. 92(3). 420–440. https://doi: 10.2174/138920312803582960.

Johnson, Deborah G. 2006. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*. 8. 195–204. https://doi: 10.1017/CBO9780511978036.012.

Johnson, Deborah G. 2015. Technology with No Human Responsibility? *Journal of Business Ethics*. 127(4). 707–715. https://doi: 10.1007/s.

Johnson, Deborah G. – Keith W. Miller 2008. Un-making Artificial Moral Agents. *Ethics and Information Technology*. 10(2–3). 123–133. https://doi: 10.1007/s10676-008-9174-6.

Johnson, Deborah G. – Merel Noorman 2014. Artefactual Agency and Artefactual Moral Agency. In Peter Kroes – Peter-Paul Verbeek (eds.) *The Moral Status of Technical Artefacts*. New York/NY, Springer. 143–158. https://doi: 10.1007/978-94-007-7914-3.

Johnson, Deborah G. – Thomas M. Powers 2005. Computer systems and responsibility: A Normative Look at Technological Complexity. *Ethics and Information Technology*. 7(2). 99–107. https://doi: 10.1007/s10676-005-4585-0.

Johnson, Deborah G. – Thomas M. Powers 2008. Computers as Surrogate Agents. In Jeroen Van Den Hoven – John Weckert (eds.) *Information Technology and Moral Philosophy*. Cambridge, Cambridge University Press.

Kane, Robert 1996. *The Significance of Free Will*. New York/NY, Oxford University Press.

Müller, Vincent C. 2014. Autonomous Killer Robots are Probably Good News. *Frontiers in Artificial Intelligence and Applications*. 273. 297–305.
https://doi: 10.3233/978-1-61499-480-0-297.

Noorman, Merel – Deborah G. Johnson 2014. Negotiating Autonomy and Responsibility in Military Robots. *Ethics and Information Technology*. 16(1).
https://doi: 10.1007/s10676-013-9335-0.

Nyholm, Sven 2017. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*. Springer Netherlands. 1–19.
https://doi: 10.1007/s11948-017-9943-x.

O'Connor, Timothy – Christopher Franklin 2019. Free Will. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward N. Zalta.
https://plato.stanford.edu/entries/freewill/.

Pereboom, Derk 2003. Living Without Free Will. *Philosophy and Phenomenological Research*. 67(2). 494–497.

Pereboom, Derk 2014. *Free Will, Angency, and the Meaning of Life*. New York/NY, Oxford University Press

Schlosser, Markus 2015. Agency. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward N. Zalta.
https://plato.stanford.edu/archives/fall2015/entries/agency/.

Sparrow, Robert 2007. Killer Robots. *Journal of Applied Philosophy*. 24(1). 62–78.
https://doi: 10.1111/j.1468-5930.2007.00346.x.

Talbert, Matthew 2019. Moral Responsibility. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward. N. Zalta.
https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

Van de Voort, Marlies – Wolter Pieters – Luca Consoli 2015. Refining the Ethics of Computer-Made Decisions: A Classification of Moral Mediation by Ubiquitous Machines. *Ethics and Information Technology*. 17(1). 41–56.
https//doi: 10.1007/s10676-015-9360-2.

Wegner, Daniel M. 2002. *Illusion of Conscious Will*. London, Bradford Books.
https://doi: 10.1073/pnas.0703993104.