Journal of Philosophy of Life Vol.13, No.1 (January 2023):64-96

Why Autonomous Agents Should Not Be Built for War

István Zoltán Zárdai*

1. Introduction

Several institutions are working on creating autonomous agents (the term covers both AIs and AIs controlling robots, and it will be shortened as AAs in the rest of the paper) for purposes of waging war.[1] AAs differ from usual automated machines because instead of their behavior, they are defined by the objectives they pursue. Modern AAs can learn and optimise their behavior to pursue their goal effectively in the given environment. Therefore, it is often difficult, even for their designers, to predict what behavior an AA will exhibit. This is not a real problem in the case of sophisticated industrial robots, or AIs trained to beat chess masters, since their behavior is geared towards achieving a handful of repetitive goals, so their behavior patterns are fixed, and the means at their disposal are limited. This paper addresses worries concerning genuinely autonomous AAs, especially ones that are supposed to make decisions in morally relevant situations. The paper argues that such AAs need to have sophisticated capacities, in some cases even human-like capacities, and this leads to specific risks that do not exist in the case of other machines.

AAs have the potential to be useful for military purposes, and there are some considerations that might make it tempting to deploy them in combat theatres: fewer human lives have to be sacrificed if only machines are destroyed; their deployment can deter war, since they might make engagement much more costly for enemies; and several tasks can be performed more efficiently by AAs because they are not subject to mood swings, are precise, do not become tired or lose focus,

* Visiting researcher, Department of Ethics, Faculty of Letters, Keio University. Office 20404 South Building Keio University, 2-15-45, Mita, Minato-ku, Tokyo 1088345, Japan. Email: zizistv[a]gmail.com

[1] Unmanned combat aerial vehicles (UCAVs) are one type of such agents. At the moment, most of the models deployed are under real-time human control or monitored by human controllers, but more and more autonomous models are being developed in China, India, Israel, Italy, Pakistan, Russia, Turkey, the USA, and other states. This development goes hand in hand with research by these and other states on increasingly sophisticated military AIs.

do not disobey orders, and are not afraid. AAs that complement existing military capacities and operate under constant human control and surveillance are already operative.[2] Such systems at the moment are not allowed to make lethal decisions; that is, they are not autonomous in the US military's sense that once they are activated, they can select and "engage" (i.e., intentionally fire at and kill) targets without further intervention by a human operator.[3]

At the same time, it is safe to assume that there are some developers and governments that agree that AAs should also be allowed to reason, decide, and act independently, without having to always involve human operators when deciding about a potentially lethal attack or an attack that is intended to be lethal.[4] If nothing else, the current trend of AA development makes this highly likely: there is no unilaterally accepted ban or constraint on the development of military AI and other robotics that would prevent this, and states are investing openly more and more into such research and encourage the integration of the latest technology into military tools.[5] This naturally leads to an "arms race" situation, as described by Chan (2019).

Those developing lethal autonomous weapon systems do not mean "autonomy" to imply that AAs should be able to choose their strategic objectives, switch which side they fight on, decide on what mission objectives they pursue, and so on. They mean that the AAs should be allowed to kill, or to use a popular military and marketing euphemism, to "engage targets" and "use lethal force." Such autonomous AAs, such as turrets, submarines, or bombers, could decide on their own when to fire missiles or torpedoes or to drop bombs.

I argue that the development and deployment of such AAs should be strictly prohibited because 1) if these systems do not possess very similar capacities to

[2] It is highly likely that the Israeli Defence Forces used UCAVs during several operations that killed civilians unlawfully, and the same is true of the US, while China has been reported to use such technology for massive-scale surveillance of its own citizens.

[3] Concerns have been raised that this definition is murky and has not been observed very closely when applying regulating procedures (see Gubrud 2015). It is known that there are deployed and currently developed weapon systems capable of performing various manoeuvring and offensive tasks without human control and oversight. For examples, see Fryer-Biggs (2019).

[4] Noel Sharkey (2010) discusses some plans, and the ethical worries about them, by the US military to transition from man-in-the-loop to full-autonomy systems.

[5] While no one admits that they want to develop "killer robots," research into offensive lethal autonomous weapon systems is known to be ongoing. Several AI researchers have lobbied the United Nations to help institute a ban on such developments, and the UN has discussed this, without much success or support from the major military technology-developing countries so far (see Glaser 2016).

humans, they are not able to make moral judgments in a way similar to humans and as such should not be allowed to act autonomously on a battlefield, and 2) in case they can be developed so as to possess very similar capacities to humans, it would be morally impermissible to create them for the sole purpose of sending them to war. In the former case, it would be unclear who is responsible for the ensuing deaths, and the AAs are not proper subjects for being held responsible and for being subjected to punishment.[6] In the latter case, we would create something that has the concept of value and can apply it in ways that humans can, and then send it out to kill and to get killed. Doing this to beings that we created, and which have capacities sufficiently similar to ours, would be unallowably cruel.[7]

In this essay, I first offer a brief discussion of what it means to be an agent and then of the capacities that ground moral responsibility in section 2. In section 3, I argue that AAs should not be allowed to act autonomously—in the above-mentioned sense, that once they are activated, they can select, fire at, and kill targets without further intervention by a human operator—because they are not the right kind of agents, since they cannot be held responsible. In section 4, I defend the claim that while human-like AAs could be held morally responsible, for moral reasons it cannot be allowed to create them. Finally, in section 5, I anticipate and address a number of possible objections, before summarizing the key points and concluding the paper.

[6] For an argument against the deployment of autonomous weapon systems on the grounds that the relations of responsibility between commanders, manufacturers, controllers, and other involved parties cannot be made sufficiently clear to warrant deploying such robots in war, see Sparrow (2007).

[7] There are two further important issues that will not be discussed here: 1) Sometimes it is argued that decision making about killing in war is somehow more mechanical or more straightforward than decision making about killing in other situations, and hence it is more permissible or even desirable to put AAs in charge of decisions than putting them in charge of important decisions in, say, politics or healthcare. There is, however, no good reason to suppose that decisions about killing are really much simpler in a war context than in other situations. They are surely of a different kind for several reasons. For example, they are unlike conflicts between civilians; they are in the contexts of political goals of communities like nations, groups, parties, countries, or alliances; they fall under specific regulations; etc. These differences do not mean that they are simpler, only that they are different. 2) Some philosophers have argued that killing in war can be understood analogously to the ethics of cases of self-defense in civilian life. This position is deeply oversimplifying and leads to morally troubling results, but since it is a reasoned and detailed position that deserves its own treatment, I will not attempt to address it here.

2. Autonomous Reasoning, Decision Making, and Acting

Agential capacities

I present in this section a broad view of agency. This view recognises artificial agents, such as AAs, as capable of acting. Then I argue that agents need to possess and exercise specific capacities in the right way. Beings that do not possess these capacities cannot be understood as choosing or as making decisions, and they cannot be held morally responsible. Hence, if AAs are to make life-and-death decisions, they need to possess these capacities. If this can be shown, the conclusion has both ethical and practical consequences. The ethical consequence is that there is a strong normative reason for not creating such AAs because it is immoral to do so, since it is like creating humans solely for the purposes of war. The practical consequence is that research aimed at creating such AAs might be reconfigured and the enormous funds that militaries, governments, and companies would spend on this could be spent on more useful research and practical activities vital for our societies.

My focus in general will be mostly on showing which capacities AAs would need to have in order to be human-like, and I will say less about why it would be wrong to manufacture such AAs for the purposes of war. This is so for two reasons: First, I think not enough has been published on the issue of exactly what kind of agents AAs are, should be, or could be. The literature on them has only a few discussions on how they compare to other agents.[8] At the same time, many arguments are available in the ethics literature that argue against manufacturing human-like AAs,[9] as well as broader ethical arguments that could be applied by analogy, such as from the literature on cloning, or Kantian arguments that could be employed to show that human-like agents are persons, and as persons they should be respected and not treated as means for ends like fighting wars.

Reasoning, decision making, acting, and moral responsibility are things an agent needs to possess and exhibit in order to act intentionally and/or voluntarily

[8] Some of the more detailed and higher-quality discussions are by Purves, Jenkins, and Strawser (2015) and Misselhorn (2018).

[9] For an argument that it is most likely impossible to programme AI to make moral decisions reliably, see Jenkins and Purves (2016).

and to be an appropriate target for being held morally accountable.[10] The role of these capacities and activities in this discussion is clear: reasoning, decision making, and intentional and/or voluntary action have to happen at some point in order to ground attributing moral responsibility to an agent and to treat it as a morally competent agent. Of course, both in everyday life and in law we hold agents responsible in cases when they are not active as agents, such as in cases of negligence (think, for example, of a security guard forgetting to lock a door) or omissions, but the cases that interest us in this essay are mainly those of active killing, and hence of intentional and/or voluntary action. Also, we only hold agents responsible who, we are more or less sure, possess such capacities.[11] This is also clearly indicated by the fact that we treat children, those living with serious mental disabilities, those in a permanent coma, and minors differently, as well as by the fact that we do not treat animals—even such highly intelligent and social species as dolphins, dogs, or chimpanzees—as responsible.[12]

Let me start the discussion by clarifying how this position relates to some of the issues that are related to but are not my main focus here. Sometimes it is claimed that if AAs have a small number of key features—like rational thinking or producing appropriate emotional reactions—in common with humans, then they are human-like. I call this way of thinking 'the smallest common denominator' approach. It is usually endorsed by people who are optimistic about the possibility of creating autonomous, human-like AAs. This does not mean that I presuppose that there is no free will or that moral responsibility and autonomy can be reduced to deterministic lower-level processes. What it means is that people who believe that creating autonomous human-like AAs is possible have to accept the premises proposed, since it is not intelligible to claim that something can be responsible if it does not meet at least these criteria (i.e., the conditions are sufficient, although one could argue that more is needed; that is, they are not

[10] It adds to the complications of creating AI that can reliably fulfil criteria of intentional and voluntary behaviour that it is an open question what exactly these concepts mean and which conception of them is the most relevant to law, morality, and war. I explore some of these difficulties in connection with interpreting the famous Knobe Effect in Zardai (2022). The literature on this effect shows what a dazzling array of potential explanations there is of how our practice of holding others responsible might work.

[11] The approach recommended here to what moral responsibility is, is compatible with a range of views of responsibility. It is particularly amenable to a Strawsonian or moral sentimentalist view. See, for example, Strawson (1963/2003: 78-81), as well as the work of Paul Russell and R. J. Wallace.

[12] Strawson 1963/2000: 78-81.

necessary). I want to show that if we accept that we can create AAs that could be autonomous and responsible, in the sense in which humans are, that possibility is an even stronger reason not to do so.

Free will

I also want to make clear that in the essay I do not discuss the free will debate, and what I say does not bear on it directly.[13] The kind of AAs that I describe as human-like are certainly determined and not free in the sense in which libertarians about free will would require them to be undetermined for them to count as free, and in virtue of having that kind of freedom, also morally responsible. Nevertheless, I think that such AAs can be morally responsible if they meet a number of conditions. My view does not claim that free will is mischaracterised by either libertarians or hard determinists about it, but it does claim that responsibility is possible without it, at least in some important cases. Hence, regarding the metaphysics of free will, my position is neutral, while regarding responsibility, it is closer to compatibilist positions. It can be accepted by libertarians and hard determinists, if they concede that something could have moral responsibility—maybe in a limited sense—even though it would not be free, or if they can plausibly claim that the moral responsibility I attribute to human-like AAs here is not the moral responsibility they are interested in because they are looking for a fuller, more demanding notion. I'm happy to accept either approach. And I think there are two good reasons why my own line is adequate. First, there are several serious and well worked out, even if not universally accepted, compatibilist positions in the literature. In this essay, I could hardly do better than those detailed discussions. At the same time, it is sensible to rely on the good work others have already done on the topic.

One of the best reasons to be a compatibilist is Harry Frankfurt's work on how to make sense of moral responsibility in cases when an agent cannot do otherwise than it does.[14] Frankfurt rejects what is called the Principle of Alternate Possibilities (PAP). The PAP says that a person is morally responsible for what he

[13] For short but excellent overviews of the major positions and debated points in the current literature on free will, see Watson (2003), Pink (2010), and chapter 1 of Mele (2017).

[14] Frankfurt (1969/1988), (1971/1988).

does only if he could have done otherwise.[15] That agents can be held responsible even in cases when they could not have done otherwise can be shown by a thought experiment proposed by Frankfurt. Imagine that Black is about to kill Jones by shooting him. Unbeknownst to Black, a team of mad scientists has planted a chip in his brain. They are monitoring Black's behavior and brain activities and are able to tell what Black wants (this is not possible at this stage of science, but it is helpful for understanding responsibility to imagine that it is). If Black for some reason decides not to shoot Jones, the mad scientists activate the chip in his brain, and this will change Black's neural activities so as to make him shoot Jones. That is, Black cannot do otherwise; he will shoot Jones. As it happens, Black sticks to intentions and, without any interference from the mad scientists, goes on to shoot Jones. He could not have done otherwise and is nevertheless responsible for shooting Jones.[16] What follows from this for the possibility of the responsibility of AI is that even if we look at AIs as machines that act according to mechanisms and are fully determined, that does not mean that they cannot be responsible for what they do. They can still be morally responsible, just as humans are.

A second reason to accept the approach proposed here is that it provides mutually agreeable grounds to libertarians and hard determinists. Since I claim that possessing the well-developed capacities that grown-up humans living without mental and emotional disabilities possess is necessary for the kind of agency that grounds moral responsibility, libertarians can say that such AAs would have free will just like humans if they would possess the same agential capacities and hence would be morally responsible. Also, hard determinists could claim that AAs lack the relevant kind of freedom, just like us, and therefore it does not make sense to attribute moral responsibility to them, the same way it does not make sense to attribute it to us. My further claim, that it would be impermissible to create free and responsible agents for the purposes of war, will still be something that they would have to address. Presumably, for libertarians, the question and how it is to be answered will not change much from cases of other free agents. For hard determinists, it will change in the same way as all other questions about how to deal with moral issues after giving up on moral responsibility are to be dealt with.

[15] Frankfurt (1969/1988: 95).

[16] For the original version of the example, see Frankfurt (1969/1988: 6-7).

To help understand my approach better, I use as a contrastive example a different, libertarian position with which I disagree: Helen Steward's thoughtful view worked out in her *A Metaphysics for Freedom*. Steward claims that

> (…) the falsity of universal determinism is a necessary condition of the possibility of any freedom or moral responsibility there might be. But the best reason for thinking that is so, in my view, is that the falsity of universal determinism is a necessary condition of *agency*—and agency is, in its turn, a necessary condition both of free action and of moral responsibility. (Steward 2012: 4)

That is, according to Steward, if something is not free, it is not an agent. This is a narrow view of what agency is—and perfectly fine for the purposes of Steward's investigations in her book—whereas I endorse a broader notion of agency, according to which several obviously non-free things are agents. Hence, for Steward, some criteria of agency will be part of the explanation of why agents are free and can be morally responsible. My approach requires the strategy that I have described earlier, namely, arguing that the kinds of agents who possess full moral responsibility—healthy adult humans—have specific capacities that distinguish them from other kinds of agents—like acids, wolves, or planets—and having these capacities is what explains why they can be morally responsible. Other agents also do things, just like humans do—acids dissolve other materials, wolves nurse their offspring, planets orbit suns—but nevertheless, they lack the capacities that would ground holding them responsible.

Consciousness

Another issue I address only tangentially is the important notion of consciousness. Every day as humans we experience the evaluative aspects of our existence: When we wake up, we feel somehow and in a kind of mood. We crave coffee, or want to go back to bed, or feel energetic and jump out from under the blanket with our minds already racing through what we need to do that day, and so on. All these cases come with their own specific experiential side that we can be aware of. That is, we have phenomenal consciousness. Due both to behavioural evidence and neurological similarities, we also have good reasons to suppose that at least the animals closest to us in their social and mental features—dogs, dolphins, elephants, primates and whales (and most mammals really)—have this "what it feels like" aspect of mental life, even if it might be qualitatively and structurally different from ours. Experts on other life forms have argued extensively that birds, and even fish, have conscious experiences.[17]

It has been argued that this experiential facet of our mental life is important for being the kind of agent we are, including having the capacities required for being responsible and for having moral value.[18] This is significant because it supports my approach that humans are the kind of agents that can be morally responsible. Hence, an AA that is human-like has to have phenomenal consciousness too in order to be eligible to be deployed in a war where it has to make decisions that only morally responsible agents should be allowed to make. At the same time, this is also why a human-like AA should not be manufactured with the aim of deploying it in a war.

For the purposes of this essay, I will work with the assumption that if an AA can be manufactured that has all the other capacities necessary for being the kind of agent that can be morally responsible, then it also has phenomenal consciousness. The reason why it is safe to work with this assumption is that several of the distinctively human capacities involve experiential content in a range of ways, including the ability to consciously reflect on that kind of content; to take it as information and grasp it both as felt and experienced, such as when we commiserate with someone or share their joy; and to handle it in a way that enables meta-reasoning (thinking about it, rejecting it, questioning it, etc.) and taking it into consideration when deciding what to do. This is so due to the key role some of the capacities, making use of such experiential information, play in making decisions that render us morally responsible.

Since we are considering what kind of agents should be allowed to act autonomously in combat situations, it is relevant to discuss a case when human

[17] Victoria Braithwaite defends the view that fish experience pain in *Do Fish Feel Pain* (2010). More recently, it has also been argued that fish can even feel "emotional fever" (see Rey et al. 2015).

[18] For the idea that robots are not agents in the relevant sense unless they are conscious, see Talbot, Jenkins, and Purves (2017). For an extended argument for the importance of phenomenal consciousness for having moral status, see Shepherd (2018). I'm not using Shepherd's arguments here but drawing on them to motivate the thought that phenomenal consciousness is relevant for having some of the core capacities involved in actions for which agents can be morally responsible.

soldiers do not act in line with prescribed deliberative procedures. The case of a US patrol in Afghanistan in Taliban territory illustrates this well. The American soldiers were nearing a dwelling known to have given shelter earlier to Taliban fighters. The locals sent out one of their younger children to look after the livestock and meanwhile also take a look at where the American soldiers are. At this point, the American soldiers had intelligence that they will likely be ambushed, and hence could have regarded the child, who repeatedly looked in their direction in an obvious manner, as a combatant because she was gathering information that could endanger the soldiers. However, the locals' gamble paid off: the soldiers refrained from shooting a child and dispersed.

The decision making of the US soldiers in this case can be reconstructed in two ways: one is to imagine that it relied on an absolute ban on killing children, and one is to imagine that it involved some form of compassion or empathy. The first is unlikely, since the rules of war posit that children can be targeted in such highly specific cases when they are contributing to the war effort and their activity constitutes imminent danger. It is also known that militaries do kill children in such situations, the US not being an exception to this. Hence, the second line of interpretation is what is more likely. The soldiers' reasoning could have taken any of a large number of imaginable avenues; nevertheless, it is plausible that they relied to some extent on emotional experiences and empathy. If compassion played a factor, then such experiential features entered the reasoning because compassion is a feeling of sorrow or pity awoken by the suffering—real or imagined—of others. If pity for the child influenced the reasoning, then information provided by phenomenal consciousness was likewise employed, since pity is an emotion of sympathy, sorrow, or regret caused by the suffering of others. If the soldiers were reminded of their own children, then emotion played a role by influencing their reasoning through invoking impressions of the sadness that the shooting of the child would cause the child's family.

Agency

Considerations regarding the importance of phenomenal consciousness led us to discuss the capacities that are important for moral responsibility. However, by doing so, I have skipped ahead. Let me say more about what is needed for something to be an agent in a wide sense. There are different views of what agency is, and what partly determines which notion of agency one wants to spell out is what work one wants the notion to do. In my case, the notion of agency serves the purpose of solving the problem of action.[19] The problem of action can be briefly introduced in the following way: The world is full of changes. Some of these changes can be attributed to agents, and such changes are actions or the results and consequences of actions. In some cases, the same change could happen to an agent while the agent is passive—say, if I fall on the bed after my dog jumps on me and pushes me over—and in an active way—say, if I fall on the bed because I'm pretending to fall after my niece shoots me with her imaginary gun while we are playing. While the changes are different in these two cases—someone's falling is not the same as someone's pretending to fall—the results occurring—my landing on the bed after suddenly leaning back—are identical. On the view of action that I defend, the distinction here can be understood in terms of the role the agent's capacities played in these two changes.[20] When I'm being pushed over and fall, none of my agential capacities are active in the sense that they are not bringing about this fall and their activity is not identical with my falling. In contrast, in the case when I pretend to fall, several of my agential capacities are actualised—my coordination, controlling my balance, and so on—and together with my body's landing on the bed, they constitute my pretend-falling (Zardai 2016, 2019).

This way of solving the problem of action has several attractions: thinking about actions as identical with or partially constituted by actualisations of agential capacities enables us to understand what actions are without relying on the concepts of reason, intention, will, plan, value, or other concepts involving the idea of higher-order mental abilities.[21] This, in turn, enables us to endorse a wide

[19] Harry Frankfurt (1978/1988) dubbed this issue so in his paper titled 'The Problem of Action.'

[20] The relevant notion of change is defined by Georg von Wright (1963) in his *Norm and Action*, chapter 3.

[21] For the most famous and influential narrow notion of agency, see the causal theory of action as presented in Donald Davidson's work, esp. Davidson (1980). The causal theory of action defines action as bodily movement caused by an intention (or a belief–desire pair that amounts to an intention), where the content of the intention is a judgment that the agent also has reason to perform the action that the movement is identical with. A current version of the causal theory is endorsed, for example, by Al Mele (2017, esp. chapters 1 and 3). I argued for abandoning the causal theory of action in detail (Zardai 2016, 2019). There are also other views of action and agency—like the new version of agent causation defended by Alvarez and Hyman—however, for the current purposes, the view I have introduced will do well in helping to capture what is needed for moral responsibility, that is, what

notion of agency. This notion of agency can be briefly presented as follows: anything is an agent that has specific capacities in virtue of its overall structure (i.e., in virtue of the relevant structure that makes it the kind of agent that it is). In the case of humans, our biological and psychological structure are both relevant to what we are and hence help highlight our agential capacities qua humans.[22] The view can also make sense of animal agency, and even of the agency of inanimate things, like chemicals, natural phenomena, and machines. As such, it works well for addressing issues about AI too, since it enables us to treat AAs as agents, without also automatically attributing them the ability to act intentionally or voluntarily, or the status of morally responsible agents. Narrower notions of agency, which define actions and agency with the help of intentionality, the will, goals, directedness, and other concepts that immediately invest a stake in normative issues, have a harder time providing an account of the behavior of robots and other artificial agents. Such agents all have a specific structure in virtue of which they are what they are, and this structure equips them with their agential capacities. According to the broad view of agency employed here, when these capacities are instantiated, the agents are acting.

So far in the essay, I have clarified the current discussion's relation to the free will debate and the role of phenomenal consciousness, and I have introduced working notions of action and agency. All this should enable us to see that in theory there is no obstacle to creating human-like AAs, since determinism allows for such agents as well as the relevant notion of moral responsibility. Also, we can propose a notion of agency and action that can explain why things like AAs are agents and how they can act. This notion of agency and action does not have any premises or presuppositions that are hard to accept: it is compatible with a broadly naturalist world view—one that accepts that science is the authority in its fields of study, while being open to the possibility of future changes and paradigm shifts in the sciences, and also being a realist about morality, values, and social entities. At this point, we can move on to discuss in which specific aspects AAs would need to be human-like to possess moral responsibility in a substantial sense. For this, we will first need to discuss what moral responsibility is and say a few words

kind of agents human-like AAs would need to be for them to make life-and-death decisions.

[22] This view of agency is inspired partly by Maria Alvarez's and John Hyman's work on action and agency. See Alvarez and Hyman (1998) and Hyman (2015, esp. chapter 2).

about autonomy.

In the introduction, I noted that the military uses “autonomous” to mean that after an agent is activated, it can select, fire at, and kill targets without further approval by a human operator. My essay is trying to show that for AAs to be allowed to do so would require AAs to be human-like. Otherwise, they would not be the right kind of agents to make a decision about the life and death of a human, or even to engage in any act of war that carries the risk of causing harm to humans, without human control and oversight. For example, if targeting would be carried out or get confirmed by a human operator, as is done currently in the case of cruise missiles when targeting other ships or planes in a battle, or the battlespace would be managed in a way that would ensure that non-combatants and allies could not be injured, the automatic operation of AAs would not be controversial. But AAs that are not human-like making decisions in such scenarios is not acceptable.

Autonomy

My point can also be understood as addressing the following issue: militaries might attempt to give different meanings to “autonomous” in order to claim that a specific machine should be allowed to operate without human oversight and control when deciding to kill. One way they might try to do this is to say that because the machine meets the required—weak—notion of autonomy, it does not need to be controlled. This strategy downplays the requirements of autonomy and tries to display many AAs as meeting them. Conversely, militaries might postulate a very strong notion of autonomy and claim that we already allow several machines to destroy targets without human intervention, such as in the case of air defense systems. Taking this strategy, militaries could say that AAs are similar to the highly automated systems that we already allow to target and kill humans without approval by a human operator. Hence, the operation of such AAs does not need to involve humans in the oversight and control of decisions to kill. My point applies to both of these attempts at loosening moral norms by watering down the notion of autonomy or making it unreasonably demanding: we need to resist such redefinitions of autonomy. Unless the specific machines in question, the AAs, can make decisions in a human-like way, *in the same way* a human operator would do, they should not operate without human oversight when making decisions about taking potentially lethal actions. Hence, just because automatic missile defense systems or drones might currently be allowed to select their targets without the control of a human operator in some cases, that does not mean that they are not autonomous in the sense of autonomous that calls for control if they target someone for killing.[23]

According to the military's technical definition, something is autonomous if "after an agent is activated, it can select and fire at and kill targets without further intervention by a human operator." This shows that there are clear technical uses of the term. These technical notions might serve useful purposes when classifying weapon systems from a quality control, testing, budgeting, or safety perspective. However, they are irrelevant when we are trying to decide whether the construction of military AAs that could kill without explicit, real-time human orders should be allowed. It is irrelevant because such technical notions of autonomy can be met by several animals and also by automatic equipment like industrial assembly-line robots or statistical AIs analysing complex data sets and creating action recommendations for humans. Still, neither animals nor such robots qualify as responsible agents. The everyday moral notion of autonomy should override the technical notion of autonomy proposed by the military, since the actions of weapon systems affect everyone and have a moral dimension. When AAs perform an action that leads to harm or other negative consequences, the ones bearing responsibility are the humans interacting with these animals and robots and the producers and operators of the robots. This model should also apply to military AAs.

The notion of autonomy that is interesting for us when thinking about whether militaries should be permitted to develop, purchase, and deploy AAs is a robust notion. If an agent qualifies as autonomous according to it, then it can also qualify as morally responsible. We do not want to claim that something has autonomy in

[23] A suspected case of such over-definition of autonomy has been highlighted in the report of Drone Wars UK in their 2018 on the development of autonomous military drones (UVACs) in the UK. There is a tension between the verbal commitment of UK politicians not to fund the research, purchase, and deployment of autonomous weapon systems, and the military's repeated announcement that it is committed to go automated to a high degree. The possibility cannot be ignored that politicians and the military would attempt to define autonomous in such a demanding sense that they can claim of UVACs and other automated weapon systems, which would meet the US Army's criteria of autonomy, that they are simply automatic but not autonomous and hence can be used under more lax regulation and supervision that currently apply to automatic weapon systems. See Burt (2018: 54-55).

this sense if it cannot bear moral responsibility for its actions. This is so because something that lacks the capacities needed for bearing moral responsibility lacks the capacities for understanding the morally charged situations that it is operating in. Such agents would not understand features of key importance of the situations they would have to decide and act in. In such circumstances, we humans could not understand properly what they do, and we could not allocate moral responsibility effectively to their operators and manufacturers either, both of which are further key criteria that need to be met before any AAs can be deployed.

It seems then that the notion of autonomy is not the most helpful concept to get at the answer to the question of whether machines should be allowed to make decisions and execute potentially lethal actions without real-time human permission. It can create confusion, which might be exploited by vested interests, and its relevant requirements can be more helpfully discussed in more obviously moral terms. Autonomy is originally a political term that means that a nation can decide about its affairs without external interference, and in this way it is connected to sovereignty.[24] Politicians, militaries, and other industry use the term in a variety of technical senses and with different purposes. What I argued for here is that there is a sense of autonomy that is tied closely to moral responsibility, and this is the relevant notion to this discussion. We should resist attempts to redefine autonomy. And to properly understand the relevant sense of autonomy and find the answer to our question, what we really need to do is to think about moral responsibility.

3. Responsibility

There are different kinds of responsibility. It is plausible to distinguish at least three kinds: causal, role, and moral. An agent is causally responsible for something if it is part of the causal chain leading to that thing. For example, the typhoon can be responsible in this way for the fallen trees on the beach. Role responsibility depends on what duties an agent has in virtue of having a specific role. For example, parents in their role qua parents are responsible for the health and well-being of their children, including that they are fed and clothed appropriately, and that they attend school at least until they reach the legally

[24] Darwall (2006: 263-265).

required age of compulsory education. What we are interested in here, moral responsibility, means that an agent can be held accountable and be the target of specific normative responses. Agents—whether individual or group agents—are held responsible for particular items, such as choices, decisions, actions, omissions, results and consequences, character traits, and so on. These kinds of responsibility can come apart.[25]

The main focus in this essay is on moral responsibility, because this is the kind of responsibility that is connected to what is morally good and bad, right and wrong, and permissible or impermissible. The praise- and blameworthiness of persons is also connected to moral responsibility, since praise and blame are the relevant normative responses helping us to map the boundaries of attributing responsibility. What are the conditions for an agent to count as morally responsible? Exploring these criteria will help to show that AAs cannot meet such criteria currently, and they could only meet them if they would be sufficiently human-like. However, if they are human-like, then they should not be created and deployed to go to war.

Normally we attribute responsibility to persons. Views of responsibility work with the idea that the agent has a set of capacities and abilities, and these play, or could play, an essential role in their actions for which we judge them responsible. Such capacities and abilities include having values and preferences; being able to reason about these reflexively; being able to feel, to be vulnerable, know that they have something to lose by ceasing to exist or getting injured and incapacitated, to possess the ability to develop empathy and sympathy, to entertain plans and goals and rank these to choose between them or revise them; and a number of other capacities as well. If we judge that an agent acted in a way that its relevant capacities were or could have been involved, we can morally judge the agent.

Moral judgment usually goes together with blaming and praising. Blaming and praising can be understood as reactive attitudes, reactions that we have towards other people. Peter Strawson emphasised the importance of such attitudes

> We should think of the many different kinds of relationship which we can have with other people—as sharers of common interests; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an

[25] Fischer (2010: 309-310).

> enormous range of transactions and encounters. Then we should think in each of these connections in turn, and in others, of the kind of importance we attach to the attitudes and intentions towards us of those who stand in these relationships to us, and of the kinds of *reactive* attitudes and feelings to which we ourselves are prone. (Strawson 1963/2003: 76)

Naturally, specific attitudes are appropriate under different circumstances, and to find the key to when blame and praise are appropriate, Strawson turns to considerations about when it is appropriate to feel resentment towards someone for what they have done and act on the basis of it towards them. We can say with Jonathan Glover, who is following Strawson here, that "To say that someone is morally responsible for what he does may be to say that he can legitimately be praised or blamed if either of these responses is appropriate to the action in question."[26] Considering when resentment, and more specifically blame, is appropriate and when it is not can help us to understand which capacities and abilities an agent must have and what external conditions must hold for blaming to be appropriate.[27] This can be our guide to understanding responsibility.

Since the publication of Strawson's original work, a variety of accounts of moral responsibility have been published. It is not crucial for us to focus on any of those here, and I will make use of ideas from more than one to help get a grasp of the capacities needed for moral responsibility. Another way to make sense of morality is to claim that "For an agent to be morally responsible for an item is (…) for that item to be attributable to the agent in a way that would make it in principle justifiable to react to the agent in certain distinctive ways."[28] Being responsible means a kind of attributability in this case, and the justified reactions would consist of judgments amounting to appraisals of the agent's moral virtues and vices, or to more inclusive judgments of the agent.[29] Whether we endorse this view of moral responsibility or Strawson's view, it is a common strand that being responsible depends on whether the agent meets the criteria of attributability.

As a quick aside before moving on to discussing the criteria of attributability, I want to summarise an important distinction between moral responsibility and

[26] Glover (1970: 19).

[27] Strawson (1963/2003: 75-79).

[28] Fischer (2010: 310).

[29] Fischer (2010: 311).

autonomy. In his overview of the two notions, Fischer argues that

> moral responsibility is a necessary but not sufficient condition for autonomy. (…) In order to be an autonomous agent (…), one must be a morally responsible agent (…). But some additional features must also be present; one can be morally responsible without being autonomous. Put metaphorically, the crucial additional ingredient is: 'listening to one's own voice' or 'being guided internally'. (Fischer 2010: 312)

I have to admit that while I find it plausible to make a distinction between the two notions, I am somewhat baffled by what exactly it would mean to listen to one's own voice in this context. The notion of autonomy is, as I mentioned earlier, historically connected to the notion of *political* self-determination. On the kind of compatibilist and naturalist approach that I have taken to agents and morality, autonomy seems not to add much to the personal dimensions of freedom or responsibility that it would be worth wanting. Hence, I will put the notion of autonomy—as I suggested earlier—aside and keep focusing on moral responsibility.[30]

It has sometimes been recognised that the conditions of praise and blame might be asymmetrical.[31] I will here work with the assumption, which to me is plausible, that meriting praise is harder than earning blame. In line with this, I also claim that to elicit a reaction of praise, one must do something both intentionally and voluntarily, which involves acting for a reason, and in case one merits moral praise, one must have acted for the right moral reason.[32] Blame comes to one

[30] Not having to rely on the notion of autonomy has further benefits too: Autonomy usually implies that agents can act against their own considered best judgments. In morally charged cases, this would mean that they can act against what they think they should do and what they think would be best. Whether humans actually have this kind of autonomy is debated (some find it too demanding a description, which humans cannot actually meet, since they cannot cut themselves loose from the motivations influencing their judgments); nevertheless, it could be possible to have AAs that can do so. Since such decisions are often irrational and go against the results of reasoning that the agent reached by using its capacities aimed at coherent, rational conclusions in line with their evaluations—emotions, desires, values—too, it seems exceedingly risky to construct such agents. For a detailed discussion of whether the structure of human agency requires agents to act in line with what they deem best—that is, to act under the guise of the good—see Kieran Setiya's 'Sympathy for the Devil' (2010/2017).

[31] See, for example, Nelkin (2011).

[32] Think of Kant's example of the merchant who, if he sets prices fairly because he wants to attract more customers, is not acting in a morally good way, whereas if he is doing so out of a sense of respect

much easier: one can be blamed when one acts intentionally but involuntarily, or unintentionally but voluntarily, and even if one acted neither intentionally nor voluntarily but neglected to do something one was reasonably expected to do.

To understand what this means for the capacities required to be a proper subject for moral responsibility, I spell out a broad notion of intentional and of voluntary. An agent acted intentionally when the agent pursued a goal in acting, wanted to achieve that goal, and this wanting (motivation) structured its behavior. This loose characterisation of intentional action captures that agents act intentionally if their action expresses the motivation—be that a desire, sense of duty, plan, or any other pro-attitude—that moves the agent to act and also specifies the agent's goal in its content. It also shows that intentionality is mainly a question of the agent's psychology and its links with the action. It also indicates that intentional actions have a means–end structure, and instrumental reasoning aimed at achieving a goal by means of doing something plays a role in them. Since agents receive moral praise or blame when they do morally good or bad things, or they exemplify their virtues or vices, agents also have to be competent in perceiving the facts that are morally relevant in a given case and treat those as considerations—as reasons—for and against acting. A wide variety of facts can provide a moral reason to do something; for example, my sister's being ill might be a reason for me to visit her, the political system's corruptness might be a reason for citizens to protest, someone's owing money to their bank might be a reason to pay it back, and so on. This is so due to the complexity of morality and requires agents who can bear moral responsibility and can judge others morally, to have the capacities and abilities to recognise all such facts and react to them adequately. In many cases, such as when my sister is ill, this will clearly require more than merely rational understanding that in such cases the right thing is to help her. Visiting my sister out of sheer adherence to a rule is problematic: I might be doing the right deed, but my doing so can hardly count as good, and doing so only out of duty highlights a questionable disposition. A certain level of emotional sensitivity and sophistication is needed to get by even in everyday life, and a great amount of experience and good judgment is required when it comes to public affairs. Serving public organisations like universities, NGOs, ministries, corporations, or political parties is a real test of one's capacities. Still, even

and duty, he is acting in such a way and deserves moral praise.

decisions made in such environments might be simpler than decisions made in war.

Regarding voluntariness, we can say that an agent acts voluntarily when the agent is free from duress and coercion while deciding to act and acting and the agent does not act in ignorance of what they are doing.[33] This notion of voluntariness captures that voluntariness is primarily defined negatively and requires freedom from specific types of social, political, violent, and other pressures and that the agent has to be knowledgeable about what they do to some extent. It also highlights that—differently from intention—voluntariness is partly about external factors rather than the agent's psychology, and nevertheless it requires that the agent have an adequate psychological life and abilities. Agents need to be able to understand whether they are subject to the relevant forms of external pressures negating voluntariness, and they have to be able to suspect that they might lack important information in a given situation to act rightly.[34] This notion of voluntariness highlights that agents need to be able to recognise their own interests, what they want, and when external pressure is applied to them, going against their own preferences. They also need to be able to gather knowledge regarding information that might be relevant and important in the contexts in which they act. This is why people receive training when they start a new job and why in several countries couples have to attend preparatory sessions when they are expecting a child.

Based on the discussion so far, AAs can only be regarded as acting intentionally and/or voluntarily if they have the right kind of perceptual, motivational, emotional, and reasoning capacities and these play a role in their activities. Certainly, AAs can be ignorant (lack relevant information). Beings that can act intentionally and voluntarily need to have capacities involved in recognising, choosing and revising goals, weighing up competing needs of others,

[33] Cf. Hyman (2015: 77). Note the similarity of this notion of the voluntary to Aristotle's, worked out in the *Nichomachean Ethics* III 1, 5, and V 8, and the *Eudaimonian Ethics* II 6-9. As Ursula Coope (2010: 439) summarises it, according to Aristotle, "1 an action is not voluntary if it is forced (1110a 1ff.); and 2 an action is not voluntary if it is done in ignorance of the particular circumstances of the action (1110b18ff)." However, this issue is tricky, since, as David Charles (2017: 13-4) argues, in many respects Aristotle's usage of voluntary resembles most closely contemporary ideas of intentionality.

[34] The characterisations I offer here capture the main aspects highlighted by Elizabeth Anscombe (1963), Davidson (1980), and Hyman (2015). Contra Hyman, I think that voluntariness is not entirely dependent on external circumstances. For his notion, see Hyman (2015, chapters 3-7).

reconciling clashing views, recognising sinister interests, and reasoning about how to achieve their goals. They also need capacities to have values, desire things, deliberate, plan, resolve inner conflicts, reflect on their own reasons, recognise facts as reasons, exercise empathy, and care for things.[35] They need to be able to know when to take the needs of others into account. At the moment, there exist no AAs with such abilities. Even the most intelligent AIs today are intelligent mostly in the sense of excelling at some domain-specific tasks in which they can rely on large amounts of energy, good-quality and precisely parsed data, and expert setup by engineers, programmers, mathematicians, and others. Nevertheless, there does not seem to be any conceptual or theoretical obstacle that would render it impossible that AAs having such capacities can be designed and deployed one day.

In order to be truly human-like, AAs would also need to understand praise and blame.[36] They must not only be appropriate targets for praise and blame, but they must be such that they themselves can blame and praise others. For this, they have to have participatory reactive attitudes themselves.[37] What does it take to blame someone? George Sher offers a helpful elucidation

> Given that anger, hostile gestures, reprimands, and the rest are so often associated with blame, we may reasonably suppose that anyone who blames someone must at least be *disposed* to react to him in each of these ways. This raises the possibility that what blaming someone adds to believing that he has acted badly or is a bad person may simply be the presence of the corresponding dispositions. (…) each such disposition is explicable in light of a single type of desire-belief pair—a pair whose components are, first, the familiar belief that the person in question has acted badly or has a bad character, but also, second, a corresponding desire that that person *not* have

[35] I'm not arguing here that such AAs would be able to solve the symbol-grounding problem. However, I assume that an AA that is human-like to the degree required to possess all the capacities and abilities needed to be morally responsible would likely be able to tackle this problem too. On some attempts to deal with the symbol-grounding problem, see Kukita (2014).

[36] Paul Russel argued that it is necessary to understand the moral sentiments and reactive attitudes, while Dana Nelkin has endorsed a similar but weaker approach to this issue, proposing that emotional capacities need to inform our reasoning, but in some rare or exceptional cases even without this it might be possible to understand responsibility. See Russell (2004) and Nelkin (2011, esp. chapter 2).

[37] Strawson (1963/2003: 81, section VI).

> acted badly or *not* have a bad character. (Sher 2006: 14-15)

Sher's account is not universally accepted, but at the moment I'm not discussing the precise nature of blame; I'm looking for clues as to what kind of capacities it involves. Sher's view gives us some interesting clues, and while he offers a reductive analysis of the dispositions involved in blaming in terms of a desire and a belief, these attitudes themselves are such that having them requires further capacities. It requires the agent blaming someone else to correctly identify that person as appropriate for attributing blame to them, appropriate for holding them morally responsible, and this involves that one can identify that the other person has the relevant capacities. It also requires a good understanding and evaluation of the given situation, good moral judgment, which, in turn, relies on knowledge of values and, arguably, also on caring for them. This is not an exhaustive list of all the capacities that Sher's view requires the blaming agent to possess. What this list does is indicate that an AA that can bear moral responsibility will possess a large number of the core capacities and abilities that underlie our social life. Such AAs would then most likely be able to become embedded in social life and take a role in it.

4. The Problem of Too Close a Resemblance

So far, I have argued that AAs can only be agents, act, and be held morally responsible if they possess the right capacities. They need to be such that they can act intentionally, voluntarily, understand blame and praise, and themselves be blamed and praised. If an AA met all these criteria, it would be human-like. I have furthermore stipulated that in case we would like to have autonomous weapon systems that can operate without human operators managing them during targeting and shooting at targets, we need to construct AAs to be able to bear moral responsibility.

This line of thought highlights a problem for advocates of AAs for war. The problem is that there are no good reasons to create a being substantially similar to humans solely for the purpose of then commanding it to go to war. Since such AAs would have the capacities of valuing, caring, empathy, conceptual thought, and reflexivity and would be able to feel and have phenomenal consciousness, it would be wrong to ignore their pains, suffering, and fear and to send them into extremely dangerous situations that often lead to death.

A connected concern is that while human soldiers can have several purposes in life—earning money to support their family, pursuing an ideal, serving their country, funding a business after their service, or pursuing their hobby—AAs that were created solely to be deployed in war would not have goals like these. To manufacture them and then sentence them to such a barren one-dimensional life would, again, be hard to justify. Furthermore, the high level of resemblance of AAs and humans would lead to the replication of some of the typical problems of waging war. If the psychology of AAs were morally responsible, then just like humans do, they might face issues of akrasia, moodiness, boredom, fear, obsessions, anger, urges to revolt or flee, temptation, corruption, and other debilitating, performance-lowering problems.

An example of how small differences in the volume of information handled and in the methods of processing that can make changes to the overall behavior of the system comes from the neuroscience of memory.[38] There have been experiments trying to supplement damaged areas of the brain that are responsible for creating memories with artificial parts to restore the capacity of humans to properly form long-term memories. Recently, this technology has been tested and shown to also improve the memory of participants by 37% in recall tests. One of the exciting aspects of thinking about such cases is that they highlight the riddles of information handling and volume. The right amount and coding of information are important for the workings of humans qua the kind of beings that they are. Otherwise, the memories will not be the kind of information that memories normally are or no performance improvement will take place. What we can take away from this for the case at hand is that if we just expect the outer behavior of AAs to be similar to human behavior, without paying attention to how the behavior was produced, then we might be ignoring the fact that there is no reasoning process behind their behavior—a very different process from our reasoning produces their behaviour—and they lack adequate grounds for being responsible. This is a problem because it cannot be guaranteed that different systems from ours will continue to behave like morally responsible agents normally do. One of the remarkable things about human behavior is the mix of

[38] Hampson 2018.

fixed and changeable dimensions of behavioural patterns and abilities. In learning and in becoming skilled at something—be that at painting, cooking, or fighting—human learning follows fixed patterns. This is why schools, trainers, and cooking schools work well. At the same time, individual creativity and ideas can give unique twists and introduce new methods to doing things. This combination of flexibility and capacity for improvement, and of fixed patterns and behaviours, is something that we all know well from our own lives. All our social practices are informed by these aspects of our being, including law, education, and caring practices like child rearing or caring for our parents when they are old. If the internal functioning of AAs is different to such a degree that they learn and adapt in different ways, that might again create problems for treating them as morally responsible. And even quite small differences might cause big divergences in behavior.

Building tools that process information more efficiently likely also means processing information in a different way than humans usually do. Robert Hampson's experiment was successful partly because he and his team could tailor the functioning of the neural prosthetic to the neural activity patterns of the subjects. They did not change the way the neural system of subjects worked; they simply helped to enhance the already existing processes. By analogy, if human-like AAs would be deployed autonomously in wars, they would not have a substantial edge over human fighters simply in virtue of being artificial. To actually make them more efficient would involve creating different information-processing modes from those that we rely on. AAs could have an advantage in how efficiently their capacities and abilities work, but their performance would still be in the human range.

We will almost certainly be able to develop in the future AAs that outperform humans in a wide range of activities. At the same time, if such AAs are not human-like, then we might simply not understand what they are doing and why. Such creatures might function in significantly different ways from us. Letting such creatures wage war is an unacceptable idea, since they cannot be held accountable. We cannot explain their behavior, which means that we do not understand their goals and reasoning. The risk in arming such AAs that are also efficient in war is enormous.

I will briefly mention one last concern raised by creating human-like AAs for the purposes of war. Presumably, we would want to model them on the biological and psychological profiles of humans who have done well in wars. What kind of people do well in combat? In command? As a pilot? And are the people who do well usually also people who act morally well? By what standards do they do well? Do the bombing sprees of the more efficient pilots also systematically result in higher civilian casualties? Do effective sergeants do more lasting psychological damage to recruits? Are highly aggressive people better at certain types of military tasks?

Taking into account how difficult it is to pin down "do well in war" and to spell out how people doing well in wars do in society for the rest of their existence, the following emerge as the key questions: Would people with solid moral principles do well in military settings? Would virtuous agents carry out orders well? Are people who do well in the army happy? Are their well-being levels normal? It seems intuitive that it is another reason against creating human-like AAs that we should not create aggressive, morally conformist, or morally inferior agents. Also, we should not create human-like agents in case they are miserable and their quality of life is below the threshold deemed desirable. I do not offer answers to the questions I have posed in this paragraph. Most likely the human resources divisions of militaries have data on the correlations between biology, psychology, and performance. The point of posing these questions here is to highlight how many problems the human-like AAs might encounter in having to serve in the military.

One might counter that I made a shallow point, since human soldiers face the same issues, or arguably even worse, since they might have enjoyed higher life qualities during most of their previous lives than during their time serving in a war. While this is true, those humans do not end up in the military without any preliminaries leading them there, nor do most of them have to spend the rest of their lives in service. I will not touch here on the point whether it would make sense to manufacture AAs that would lead full human lives, with a period of military service. This question poses different issues and is a separate topic.

In this section, I provided some considerations against manufacturing and deploying human-like AAs for the purposes of war. I did not rehearse standard Kantian and other deontological arguments, nor the often-repeated utilitarian arguments. My goal was to tie in the points I made with the preceding discussion regarding the capacities and abilities on which agents rely when they act intentionally or voluntarily. In the next section, I will address potential objections.

5. Objections Addressed

Two main types of objections could be raised against my position. Some objections try to show that it is actually good if the AAs deployed in war are different from humans. Other objections try to deny that human-like AAs are possible and claim that arguing against their possibility is not important. I reject both of these types of argument. I think the first type of argument gets something right: AAs that are not like humans can be beneficially employed in many roles in wars. But they cannot be allowed to make substantial morally charged decisions, including decisions to cause harm without the real-time management of a human. The second type of objection relies in most cases on too-demanding ideas about what counts as a morally responsible agent. Maybe some libertarians would claim that unless one can have a capacity for self-determination and hence for freedom, one cannot have moral responsibility. I addressed such worries in section 2. I would also want to add at this point that there are several convincing views of moral responsibility that do not depend on libertarian ideas. I tried to give a rough sketch of the capacities and abilities AAs need to count as morally responsible. If I was moderately successful, then the worry that they could be misused for purposes of war in morally bad ways is real.

Another objection would go like this: Using AAs in war has an advantage over deploying human soldiers exactly because AAs reason differently. Military law does not treat soldiers as persons in the same way we normally treat civilians as persons: soldiers are *not* supposed to make decisions on their own, follow their own values consistently, and so on. A good soldier—especially a lower ranking one—should follow orders. Employing humans as soldiers poses all kinds of issues: there are several historical examples of having to force soldiers to attack, soldiers fragging their officers to avoid having to follow them into a charge, and so on. In this sense, if AAs do not have normal reasoning processes, that is an advantage. They can follow orders more easily and more efficiently, without this causing them emotional distress and harm, and without causing trouble for their forces.

I do not claim that denying some autonomy for AAs is bad. As I said, AAs that are deployed for well-described tasks of limited scope—and if that scope includes potentially lethal engagements that are managed in real time by human operators—can be of much use to any military. What I claimed was that a human-like agential constitution—in line with the view of agency introduced in section 2—would be needed for AAs to be such that we can allow them to make decisions regarding killing. Support AAs and other systems could be fruitfully developed and deployed without exemplifying substantial resemblance to human psychology in case they do not take any direct harmful actions against humans.

Another objection would say that the concept of a person is tied to the notion of *birth* and through it to the notion of *life*. Something not born cannot hence be a person, and therefore AAs cannot be the kind of thing that would be morally autonomous anyway. So, no matter what their constitution is or how they function, they can never be allowed to make decisions about life and death. A different version of this objection would give up on the idea that AAs that can make decisions about life and death have to be exactly like human agents in their origin. It would instead say that the specific hormonal processes, childhood experiences, memories, and so on that humans typically have while growing up all play a fundamental role in our becoming persons, that is, becoming the kind of morally competent agents who can be allowed to make decisions about killing in war. The objection would go on claiming that something that has a different background—biological, psychological, social, cultural—cannot be a person, and as such, it cannot be a being that reasons sufficiently similarly to us to be allowed to make weighty moral decisions. Hence, moral responsibility cannot be attributed to AAs even if they have human-like capacities.

The position I advocated here can reply that what is demanded for something to be morally responsible is not that it should be exactly like a human person, including the complete historical background of typical healthy adult humans. Rather, what is demanded is that the reasoning and decision-making processes of such agents be like that of humans. This does not require them to be persons in the sense of having a typical human birth and childhood. Being human requires more than being morally responsible. For example, if we would want to understand what human persons are, we would need to go beyond moral responsibility and also offer an account of their autonomy. But as I have argued, this is not essential for military AAs.

A further objection that could be made is that particular processes of reasoning, such as that of weighing options like when to kill in a combat situation, can be programmed well without having to provide an AA with a set of values and the ability to grasp that those are its values, and without the ability to recognise facts as moral reasons or to rationally reason to conclusions while taking emotional and moral facts into account. AAs that do not do these things—one could think of the automated machine gun turrets of the South Korean military in the border zone with North Korea—can actually be more efficient than humans in choosing whom to target. They are not swayed by emotion and other biases that humans are subject to.

To deflect this objection, it can be noted that it is unlikely that without representing the information that certain values are their own and that such values can be shared by other beings, AAs cannot treat values in a way similar to our reasoning processes. This makes it clear that they are not morally competent agents who are also more efficient than us. Rather, they lack any grasp of values. What guides their behavior then? Pre-programmed preferences and goals, which they cannot reflect on, unlike humans. I argued in section 4 that deploying agents that have non-human reasoning and permission to act without real-time management by humans is inherently dangerous. Taking such a risk is not justifiable.

There might be some exceptions, namely those cases in which a machine is reasoning in a non-human way, but its reasoning—whom it targets, when it fires, etc.—is clearly and completely understood by its operators, and it can be explained to others—law enforcement, international tribunals, expert committees—in terms that make the responsibility of its operators clear. This is easiest to achieve when such military systems are simple and set up with accountability in mind. My paper says nothing against the deployment of such weapon systems. Criteria of their ethical operation depend on the transparency of their operation and the precise calibration of their range of deployment. Such machines will inevitably be able to perform only much more primitive tasks than any human. AAs and other machines that are complex and at the same time reason in different ways from humans do not meet such criteria.

In this section, I answered a number of anticipated objections and questions that the position I argued for could face. This concludes my discussion of the topic at hand, which is whether it is useful to develop robots that can kill in war without human supervision and management. The conclusion of my paper is that it is not permissible to do so. The reason is simply that such robots would need to be too human-like to make it permissible for them to operate without constant human supervision. Otherwise, they could not have moral responsibility, and nothing lacking moral responsibility should be created to make decisions about killing. I argue for this by showing that to be morally responsible—the right kind of agent to make life-and-death decisions—requires the possession of several high-level capacities. Nothing that is dissimilar from humans—more or less complex, or functioning substantially differently—has this status. The richness of the agential capacities and abilities needed to be morally responsible is such that it makes AAs that meet this criterion too human-like to manufacture for the sole purpose of deploying them in war. Doing so would amount to creating slaves doomed to miserable lives, which is clearly impermissible.

6. Conclusions

For AAs to be candidates for morally responsible agency in combat situations, they would need to resemble the functioning of healthy adult humans in great many respects than they currently do or as is a morally good thing to grant them. The deployment in wars of AAs that would be sufficiently human-like to be allowed to make autonomous choices, without the involvement of human supervision, would pose several issues for law and morality. Human rights issues would emerge, and their human-like functioning could lead to psychological and performance problems too, in the same way it does in the case of human soldiers deployed in war. Hence, their deployment might offer no moral or other benefit over deploying humans, while producing downsides, like the questionable employment of human-like beings for the sole purpose of war.

At the same time, we could not attribute moral responsibility and hence allow AAs that are not human-like to make decisions about killing if they are not relevantly similar to human agents, except for some simple agents, the workings of which can be fully grasped, operating in strictly restricted domains. Whether it would make sense to create and deploy such AAs for offensive purposes in wars is not clear at all.

At the same time, AAs that would reason like humans would not do substantially better than humans on specialist tasks. This would undermine the purpose of using them in the first place. AAs are best not used at all without real-time supervision by humans in cases when moral decisions concerning harm and killing are involved. Where AAs can be made best use of is in well-defined, specific roles, where their dissimilarity from humans can be an advantage and can be utilized fully. AAs should then be mostly complementary systems for humans, compensating for their weaknesses. There are several roles in militaries where such AAs could be made good use of, such as in reconnaissance, where being able not to move and to tirelessly observe can be an enormous boon; in sweeping minefields, where no complex reasoning is needed; as medical assistants that are not vulnerable and might be able to carry large amounts of supplies and approach targets under fire; or in supplementary data analysis and advisory systems. If any uses of AAs would be encouraged, they should be that of non-lethal support AAs.

* The author is thankful for the support and funding he received during his tenure as a JSPS Postdoctoral Fellow at Keio University, Tokyo, Japan (ID No. P17783), and as a contributor to Dr Tamas Demeter's project "Values in Modern Science: Epistemic, Social, Moral" at the Hungarian Academy of Sciences.

References

Alvarez, M. and Hyman, J. (1998). 'Agents and their Actions.' *Philosophy* 73 (2): 219-245.

Anscombe, G. E. M. (1963). *Intention*. 2nd ed. Oxford: Blackwell.

Braithwaite, V. (2010). *Do Fish Feel Pain?* New York: Oxford University Press.

Burt, P. (2018). *Off the Leash. The Development of Autonomous Military Drones in the UK*. Drone Wars UK. Accessed: 13/11/2019. https://dronewarsuk.files.wordpress.com/2018/11/dw-leash-web.pdf

Chan, M. K. (2019). 'China and the U.S Are Fighting a Major Battle Over Killer Robots and the Future of AI.' *Time*. Accessed: 13/11/2019. https://time.com/5673240/china-killer-robots-weapons/

Charles, D. (2017). ‘Aristotle on Agency.’ *Oxford Handbooks Online*. Online publication date: May 2017. Accessed: 14/11/2019. DOI: 10.1093/oxfordhb/9780199935314.013.6

Coope, U. (2010). ‘Aristotle.’ In T. O’Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Darwall, S. (2006). ‘The Value of Autonomy.’ *Ethics* 116: 263-284.

Davidson, D. (1980). *Action and Events*. Oxford: Oxford Clarendon Press.

Fischer, J. M. (2010). ‘Responsibility and Autonomy.’ In T. O’Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Frankfurt, H. (1969/1988). ‘Alternate Possibilities and Moral Responsibility.’ In H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Frankfurt, H. (1971/1988). ‘Freedom of the Will and the Concept of a Person.’ In H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Frankfurt, H. (1978/1988). ‘The Problem of Action.’ H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Fryer-Biggs, Z. (2019). ‘Coming Soon to a Battlefield: Robots That Can Kill.’ *The Atlantic*, published on: 03/09/2019. Accessed: 13/11/2019. https://www.realclearpolitics.com/2019/09/03/coming_soon_to_a_battlefield_robots_that_can_kill_485045.html

Glaser, A. (2016). ‘The UN has decided to tackle the issue of killer robots in 2017.’ *Vox*, published on 16/12/2016. Accessed: 13/11/2019. https://www.vox.com/2016/12/16/13988458/un-killer-robots-elon-musk-wozniak-hawking-ban

Glover, J. (1970). *Responsibility*. London: Routledge and Kegan Paul.

Gubrud, M. (2015). ‘Semi-autonomous and on Their Own: Killer Robots in Plato’s Cave.’ *Bulletin of the Atomic Scientists*, published on 12/04/2015. Accessed: 13/11/2019. https://thebulletin.org/2015/04/semi-autonomous-and-on-their-own-killer-robots-in-platos-cave/

Hampson, R. E. (2018). ‘Developing a Hippocampal Neural Prosthetic to Facilitate Human Memory Encoding and Recall.’ *Journal of Neural Engineering* 15: 036014. DOI: 10.1088/1741-2552/aaaed7

Hyman, J. (2015). *Action, Knowledge, and Will.* Oxford: Oxford University Press.

Jenkins, R. and Purves, D. (2016). ‘A Dilemma for Moral Deliberation in AI.’ *International Journal of Applied Philosophy* 30 (2): 313-335.

Kukita, M. (2014). ‘Can Robots Understand Values?: Artificial Morality and Ethical Symbol Grounding.’ *Proceedings of 4th International Conference on Applied Ethics and Applied Philosophy in East Asia* Feb. 2014: 65-76.

Mele, A. R. (2017). *Aspects of Agency. Decisions, Abilities, Explanations, and Free Will*. New York: Oxford University Press.

Misselhorn, C. (2018). ‘Artificial Morality. Concepts, Issues and Challenges.’ *Social Science and Public Policy* 55: 161-169. DOI: 10.1007/s12115-018-0229-y

Nelkin, D. (2011). *Making Sense of Freedom and Responsibility*. New York: Oxford University Press.

Pink, T. (2010). ‘Free Will and Determinism.’ In T. O’Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Purves, D., Jenkins, R. and Strawser, J. B. (2015). ‘Autonomous Machines, Moral Judgment, and Acting for the Right Reasons.’ *Ethical Theory and Moral Practice* 18: 851-872. DOI: 10.1007/s10677-015-9563-y

Rey, S. et al. (2015). ‘Fish Can Show Emotional Fever: Stress-induced Hyperthermia in Zebrafish.’ *Proceedings of the Royal Society B: Biological Sciences* 282 (1819): 20152266. DOI: 10.1098/rspb.2015.2266

Russell, P. (2004). ‘Responsibility and the Condition of Moral Sense.’ *Philosophical Topics* 32: 287-305.

Setiya, K. (2010/2017). ‘Sympathy for the Devil.’ In K. Setiya. *Practical Knowledge*. New York: Oxford University Press.

Sharkey, N. (2010). ‘Saying ‘No!’ to Lethal Autonomous Targeting.’ *Journal of Military Ethics* 9 (4): 369-383.

Shepherd, J. (2018). *Consciousness and Moral Status*. New York: Routledge.

Sher, G. 2006. *In Praise of Blame*. New York: Oxford University Press.

Sparrow, R. (2007). ‘Killer Robots.’ *Journal of Applied Philosophy* 24 (1): 62-77.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Strawson, P. F. (1963/2003). ‘Freedom and Resentment.’ In G. Watson (ed.). *Free Will*. Oxford University Press.

Talbot, B., Jenkins, R. and Purves, D. (2017). ‘When Robots Should Do the Wrong Thing.’ In P. Lin, R. Jenkins and K. Abney (eds.). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.

Von Wright, G. H. (1963). *Norm and Action*. London: Routledge, and Kegan Paul.

Watson, G. (2003). ‘Introduction.’ In G. Watson (ed.). *Free Will.* New York: Oxford University Press.

Zardai, I. Z. (2016). *What Are Actions*. PhD Thesis, defended at the University of Hertfordshire. https://uhra.herts.ac.uk/handle/2299/17222

Zardai, I. Z. (2019). ‘Agents in Movement.’ *Mita Philosophical Association Journal* 143: 61-84.

Zardai, I. Z. (2022). ‘Making Sense of the Knobe-effect: Praise Demands Both Intention and Voluntariness.’ *Journal of Applied Ethics and Philosophy* 13: 11-20.