Comments for the DIB Consultation on AI Policy Principles

Associate Professor Seth Lazar, Australian National University

Introduction

Recent months have seen organisations from small corporations, to national and regional governments, and multinational tech companies, putting forward principles to guide their adoption and use of artificial intelligence systems. A number of themes have emerged, to the extent that each new set of principles tends to be a subset of the last. Most lists of principles confusingly include general goals that should be aimed at in any endeavours of a given organisation, alongside very specific issues raised by AI in particular. I welcome the DIB's approach of not simply rehashing existing policy guidelines, but taking time to think seriously about what makes AI different, and why we need new principles to govern it. I would encourage them to spend at least as much time thinking about what makes Defense different from the other organisations that have set out these policy statements. These differences, I think, make the principles around which other organisations have coalesced much less useful for Defense. I'll begin this comment by asking first what makes AI different, then what makes Defense different. I'll then ask what the goal is, of developing a set of AI policy principles. Finally I'll consider which kinds of principles we can propose. In particular, I will argue that some of the principles that so frequently appear in other organisations' lists should apply quite differently, if at all, to Defense.

What Makes AI Different? Why Does that Matter?

What is AI?

The term 'Artificial Intelligence' is exceedingly vague. We cannot hope to provide a definition of what makes something AI. So it is better to focus on those aspects of AI in which we are interested for the purposes of this endeavour. Our primary interest is in machine-learning based predictive analytics and decision technologies. We could extend this to include symbolic-logic-based decision technologies, especially insofar as they are integrated in new ways with machine learning and new capacities for data gathering. But the revolution that prompts this call for a set of AI policy principles is grounded in machine learning, and symbolic-logic-based AI has been part of military systems for a long time. Machine learning itself is dependent on data. So when thinking about policies for the use of AI, it is best to think about policies that govern not only how AI systems are used, but the preconditions for using them—in particular, how data are gathered. It's also worth remembering that the 'Gee-Whiz' approach to AI is generally unhelpful: as it stands, it is a deeply fallible technology. In some areas it offers improvements on human judgment, but it is by no means perfect.

My task in this section is to ask what makes AI, so construed, different (n.b. a difference in degree is just as relevant here as a difference in kind). Since our topic is what to do about AI, I'll focus on the differences that I think are morally important. I'll identify the difference, then say why it is morally important—in particular, by identifying the moral risks associated with it. To do this, it is helpful to distinguish between risks that arise from the nature of the technology itself, and those that arise from its effects in use. I'll talk about them in turn.

AI is a Decision Technology

At heart, AI is a decision technology. Even when its function is predictive, it is aimed at decision support, and its whole method of interpreting data is grounded in iterative decisions made within machine learning algorithms. This feature of AI is its most interesting, and most significant, difference from other technologies. To be clear, though, there are many very simple mechanistic systems that, in effect, 'make decisions'. It's hard to pick apart what distinguishes AI from, say, an elevator. It's likely to be a cluster of properties rather than any single thing. But one key feature of AI is that it is a *probabilistic* decision technology. Elevators are closed systems based on a representation of the world that has no room for uncertainty.

Why does it matter that AI is a decision technology? First, while every technology is shaped by the values of the society that makes it (and then shapes those values in turn), AI is distinctive in that those values are written into the code and are in principle revisable. In some cases they have to be explicit. Second, because of the relative success, speed, and scalability of AI, we are now seeking to embed it in seemingly every area of human endeavour. There is literally no scenario where the possibility of understanding the world better, and making better decisions under uncertainty, is not intoxicatingly attractive.

So, what difference does it make that AI is being used to make many decisions that were previously made by humans, as well as many other decisions that we didn't, before AI, have the capacity to make? Why does this difference matter morally? I can think of three reasons (here as elsewhere I draw on the rich public discussion of these topics; but since this is a public submission rather than a scholarly document, I will not attempt to trace the origins of each idea).

1. I think the key point is that when decisions are made by AI, the responsibility for those decisions is *diffused* and *diluted*. This is because (and on the assumption that) AI cannot be responsible for its 'actions'. Responsibility is diffused, in the following sense: instead of being able to attribute the decision primarily to the human who made it, as well as to any who knowingly set her on that course, the decision must instead be attributed to the humans who designed software and hardware elements of the AI system, those who developed the training data set, those who set it in motion, and those with final authority over its decisions (this is an incomplete list). Responsibility is diluted, insofar as the nature of machine-learning based AI is such that AI systems are never wholly predictable by their designers or operators, so a key precondition of moral responsibility—that one can foresee the consequences of one's action—is either unsatisfied, or only partially satisfied.

With any of these claims, it is important to bear in mind the contrast with having humans make the relevant decision. Humans are often just as unpredictable and inscrutable as AI systems. However, if one person implements a decision that is in some sense collective (for example, a combatant carrying out an order), we can hold the final person in the causal chain accountable in a way that is not true for an AI system. So if we hold everything constant besides whether the final actor is an AI system or a person, there is still a dilution and diffusion of responsibility.

Why should we care if responsibility for AI decisions is diluted or diffused? Notice that this has nothing to do with whether the AI makes the right decision or not. The success rate of the AI system is in principle completely independent of these facts about responsibility. And we might reasonably care most about the actual results of the system. However, we do not in general care only about outcomes—we care also about process. Even a guilty person deserves a fair trial, for example. And as well as wanting our decision technologies to make the right decisions, we also know that perfection is unattainable, and in the event that they get things wrong, we fundamentally want to have someone to blame. This serves two purposes. One is deterrence and guidance: if people know that they will be blamed for wrongdoing, that gives them additional motivation to do the right thing. But we also care about apportioning blame after wrongdoing appropriately, independently of these instrumental benefits. Note that this is not the same as saying that, eg, punishment is a non-instrumentally valuable response to guilt. I am making only the weaker claim that there is value in being able to blame those who act wrongly, when wrongdoing is done. If nobody is really fully blameworthy, then that deprives us of an important way in which we respond to wrongdoing. We can state this point in a general form: **the use of AI systems to make decisions threatens to undermine our practices of accountability**.

2. There is a second, somewhat more speculative reason to regret the vesting of decision-making authority in AI systems. Many people clearly have the intuition that they would prefer it if decisions affecting their lives were made by a person, rather than an automated system. It is often quite difficult to extract the rational kernel from this argument, but I think it is something like this. We are social animals. We value relations of mutual respect and care. Those relations are advanced when we make decisions that have significant impacts on one another's lives ourselves. Making decisions ourselves, rather than vesting them in automated systems, involves (when done conscientiously) paying attention to others, taking them seriously in one's deliberations.

Again, it is important to remember that the AI system is just one point in the causal chain that would lead to a particular outcome. Within that causal chain, there will undoubtedly still be human decision-makers. However, the further removed they are from the people ultimately affected by that causal chain, the less able they are to attend to them, as individuals, in their deliberations.

In a general form: **the use of AI systems in decision-making threatens to undermine the degree it which we are seriously attended to in the moral deliberations of those whose actions materially affect our lives.**

3. The third morally relevant difference arises from the distinctive nature of contemporary AI systems, and in particular the often-noted point that they are not readily understandable by either their designers or their operators. This is a technical claim, and work is currently underway to remediate this concern, for example by testing an algorithm's sensitivity to various changes in the underlying variables. Nonetheless, it's true now that if we vest decisions in AI systems grounded in machine learning, we will be able to verify the outcomes of the system's decision-making, but we cannot always explain why the system reached that decision.

Again, this merits comparison with human decision-making. We often rely on intuition and instinct in our own decision-making. It can be hard to explain our reasons for acting. And yet we can be called upon, ex post, to rationalise our behaviour. And we can be called out if our rationalisation is self-serving, insincere, or otherwise flawed. So there is a genuine contrast between algorithmic and human decision-making. But why should we care about it?

We should want people not only to do the right thing, but to do the right thing for the right reasons. This is in part because we care about their character, not only about the results of their actions. But we also care about the kind of attitude they display towards other people—and acting on the right reasons is one way to show appropriate respect for your moral equals. Another way to look at this: we want people to do the right thing not by mere luck. We want them to do the right thing robustly. If they act on the right reasons, this suggests that they would do the right thing even if the circumstances were somewhat different. Another key concern is that sometimes, some considerations might be relevant to predictive accuracy, but might be the wrong kinds of reasons to base one's decisions on (on analogy with inadmissible evidence in law, and also see the point about discrimination below). **When we vest decision-making authority in inscrutable AI systems we might make better decisions, but we may not know the reasons for which those decisions were made.**

AI Depends on Data

The collection and operationalisation of data is of course nothing new. This is a case where the difference made by AI is scale, speed, and effectiveness. Advances in AI make it possible to do incredible things with the data that we gather, and thereby incentivise gathering ever more data. This raises some obvious problems. Since they have already received considerable attention in the literatures on AI, to which I don't have much to add, I will discuss them briefly.

1. Individual privacy. Below we'll come to the ways in which AI can be *used* to undermine individual privacy. But concerns about privacy are also intrinsic to the nature of machine-learning based technology, at least when it is used for decisions about people. The simple observation here is that if an AI system is to be used to make decisions about people, it will be more effective the more data it has about them, so this creates an incentive to gather ever more data about us. Even if we retain some kind of control over that data—consenting to its use—there are real questions about whether that kind of consent is meaningful, and whether, even if we do consent, we should create a world in which the sphere of freedom in which we are not observed is ever diminishing. **AI technologies inherently threaten individual privacy**.

2. Discrimination. This is now the most readily recognised problem with the data-dependency of machine learning algorithms. Our datasets reflect structural injustices in the world as it is, and algorithms that learn from those datasets inherit those injustices. This is in fact part of a more general phenomenon—if decisions are going to be made based on historical data, then that builds in an unavoidable conservatism into our practices of decision-making. However, the key point here is: **Implementation of AI against a background of unjust social discrimination is likely to perpetuate and exacerbate that discrimination**.

The Effects of AI

The considerations just adduced all have to do with the nature of AI. But the other thing that makes AI distinctive is simply its capacity to affect every area of human life, making things possible at a scale and speed that was never possible before. AI can in principle be used for any purpose. Almost anything we can now do that is of moral concern can be done faster and at greater scale with the aid of AI. So the moral issues raised by the *use* of AI technologies cover everything that matters. This is an important point: it means that any set of principles governing the *use* of AI should really be a concise statement of the principles governing society as a whole (it also means that the principles governing AI should to a large degree be the same as those governing any general purpose technology).

There is lots of excellent research on the potential problematic effects of AI. I don't have anything really to add to it, but it might help to provide some overarching structure with which to think about them. The moral riskiness of a given application of AI seems to be a function of three factors: stakes, pervasiveness, and degree of autonomy.

By stakes, I mean: how much does this particular application matter morally? Does it significantly affect people's lives? Does it put individual rights at risks? And so on. Some use-cases are morally relatively neutral, at least on their face. Personal assistants and music recommendation algorithms, automated mining platforms and vehicles, photo editing algorithms and so on. As I'll observe below, there are ways in which even these can be morally significant—mostly insofar as they impact on what I'll call 'recognition goods'. But on the whole they involve AI systems that are not making explicitly morally-loaded decisions, so where the stakes are lower than they would be if the system were making high stakes decisions. It's worth observing that sometimes the stakes for individuals might appear to be low stakes, but for communities as a whole they are high stakes (eg when trading my data is individually rational but collectively irrational).

But even when the stakes of particular decisions are low, if a given AI system is pervasive within society, that can itself raise the moral risk. The very fact that one cannot escape it becomes an issue. If every algorithm for editing photos is much more successful at editing pictures with white faces in them, than with black faces, then that's much more of a concern than if there is enough competition, and one can avail of an alternative that suits one's needs. If I am refused credit by one bank, that might not represent a significant moral risk, but if every bank is using the same credit score algorithm, and I fall into a blind spot that they all share, then that is a big deal. And pervasiveness matters also because of vulnerability. The more pervasive a system is, the more dependent we are on it, so the more its vulnerability to attack maters for society, other things equal.

Lastly, degree of autonomy obviously raises moral risks too. By autonomy, here, I mean the ability of the AI system to affect the world without intervening confirmation or verification from a human operator. The less autonomy an AI system has, the greater the prospect there is for decisive human control, reducing the marginal difference between AI-decision-making and existing ways of making decisions. Some AI systems are really just decision-support tools. That doesn't mean they involve no risks—in particular, we need to be very cautious about automation bias, the tendency of human decision-makers to defer to automated systems. But it does reduce their risks relative to situations where they are fully autonomous.

These factors are independent of one another. The moral risks of a given application of AI might be significant if only one of them is raised. It should be quite easy to list the different ways in which AI can be used, and measure their moral risks in terms of these three factors. The use of AI to generate 'deep fakes', as well as adversarial uses of AI to spoof machine learning systems, clearly meet the high stakes criterion, though not so much the pervasiveness and autonomy criteria. Uses of AI to surveil a population are much the same, though they also threaten to be pervasive. AI for autonomous vehicles involves high moral stakes, and high autonomy, and potentially high pervasiveness. Use for medical diagnosis involves high stakes, but low pervasiveness, and hopefully low autonomy (since it is just a system for making recommendations to trained professionals). Government uses of AI for service delivery and welfare allocation threaten to be high stakes and pervasive; it is yet to be seen how autonomous they will be, though in these cases especially the risk of automation bias is very high.

AI and Power

One further gestalt effect of the rush to adopt AI technologies is worth drawing out. It is the result of a suite of different AI applications, as well as the other technologies on which they are based. AI systems have the capacity to radically alter existing power relationships between citizens and corporations, citizens and governments, between national governments and non-citizens, and between national governments.

AI enables control of information, in two directions. It controls the information that citizens receive. This shapes our view of the world, as has been much reported—enabling polarisation, and the spread of misinformation. AI also generates unprecedented information about us as individuals, making it available to both governments and corporations. This in turn enables them to shape our options, both how we spend our money, and how we vote. In my view this is a much more consequential implication of our reduced individual privacy than its implications for our ability to maintain a sphere in which we are not observed. Even if every individual whose data was used by these systems for the prediction and manipulation of behaviour consented to that use (and even if it was individually rational for them to do so) the collective implications of our 'data profligacy' would still be seriously morally objectionable.

And of course, the mere fact of automating decisions that used to be made by humans changes power relations. For example, automated performance management like what has just been revealed at Amazon means that performance management is benchmarked against context-free general standards, without sensitivity to individuals' particular circumstances. Being managed by a person with whom you have some kind of personal relationship is a very different experience from being managed by an algorithm. The same goes for the use of AI to deliver government services and allocate resources.

It's also important to be clear about the ways in which AI generates self-perpetuating power structures. It is inherently monopolistic—increasing data increases competitiveness, more competitive then more data. It's a cycle towards oligarchy. And the same issues apply to AI and international relations. The key issues here seem to be to do with cyber security, as well as deep concern about the ability of potentially unconstrained adversaries to make substantial advances in AI that we cannot match.

Any assessment of the potential impact of AI and society must stare clearly into the face of the new power structures that these systems make possible. Large multinational corporations have long been only imperfectly subject to the authority of national governments. But, through their technology, they now have considerable power directly over the citizens of those national governments. This is not entirely new. Large media corporations have long had a similar degree of influence over individuals in countries that are only able to imperfectly exercise control over them. But new technologies, AI among them, enable this to proceed at a greater scale, and efficiency.

Lastly, it is crucial to note that we should be all the more concerned about the potential impact of AI on power when we realise that AI systems are designed in large part by a very narrow demographic. Research by AI Now makes clear that the tech industry is hopelessly unrepresentative of the communities that its systems may end up governing.

What Makes Defense Different?

The foregoing identifies some of the distinctive moral risks raised by widespread adoption of AI systems. Some of these risks are genuinely social risks—risks to society, which must be considered by any organisation considering the adoption of AI, and indeed by national governments as a whole (and supranational organisations). Some are specifically risks that the developers of AI systems should be concerned with. Thus far, we have seen policy principles developed by groups on behalf of national and supranational governments, as well as standards for conduct developed by technology companies working directly with and on AI. Defense is a very different organisation from either of these different kind of groups. Most importantly, it is *very* different from the technology companies that have so far set the terms for lists of AI policy principles.

The key differences, to my mind, are in the strategic purpose of the organisation, and in the people who are ultimately affected by its decisions. The goal of Defense is to uphold the constitution of the United States, and to protect its citizens against foreign and (to a lesser extent) domestic threats. That strategic framing is quite different from what Google's goals are, eg. It's hard to say what the purpose of the tech corporations is (beyond making their owners fabulously wealthy). And it's different from eg the EU, since they have to think about all of their citizens' interests, not just those related to security. Perhaps most importantly, Defense operates in a distinct strategic environment where its very task is to anticipate and consider threats to the US. So where a corporation's AI principles might only tangentially address potential malicious uses of AI, for Defense those situations should be at the centre. AI policy principles for Defense should address not only how Defense will develop and use AI, but also how it will respond to threats that make use of AI.

Because of the different strategic purpose of the organisation, policies of the Department of Defense will affect different cohorts of people than will be affected by the other organisations. Now, of course they'll likely affect similar actual people—Google's use of AI affects US citizens, non-citizens etc. The difference is that DoD will affect people who stand in importantly different relations to it: citizens, non-citizens for whom the US has a significant duty of care, and non-citizens for whom the US has a much less substantial duty of care. These roughly correspond to domestic operations of Defense, non-domestic operations outside of a war fighting context, and non-domestic operations in a war-fighting context. As a result, DoD is subject to different legal standards—roughly, US domestic law, international humanitarian law, and the law of armed conflict, respectively.

It's also worth observing that Defense is different from the private companies that are developing AI policy principles insofar as it is much more tightly regulated than they are. It is, after all, part of the government, subject to the usual checks and balances appropriate to that station. Since it is in general subject to a much more prescriptive system of law, the first step when considering the adoption of AI systems is to ask how existing law affects them.

Finally, Defense has existing structures of hierarchical decision-making and command responsibility that are not obviously present in other organisations. In many important respects, members of the US armed services have *fewer* rights than ordinary citizens. There are kinds of discrimination that are legally permissible in the military but not in other areas of society. Service-members are not presumptively entitled to challenge decisions that affect them. And so on. But also when the military acts, it is generally not the case that the individual at the sharp end of the spear is uniquely responsible for the results. Instead, there is a structure of command responsibility which enables members of the military to function, in effect, as a kind of group agent. This will be important below.

What is the Goal of a Set of AI Policy Principles?

As I understand it, the DIB is aiming to recommend a set of principles to govern the adoption of AI technologies by all aspects of Defense, affecting all of the constituencies described above. This would cover war fighting, humanitarian and other non-war fighting overseas operations, and domestic operations including personnel, management etc. It would cover both high moral risk applications of AI, and much more benign ones, like the use of AI systems for predictive maintenance of Defense assets. The principles should obviously do more than state the obvious. The operations of DoD are obviously subject to various bodies of law—US law for domestic operations, and international law for international ones. Beyond specific legal prohibitions and prescriptions, the DoD is also bound to uphold the US constitution, which provides a guiding set of values that should shape the adoption of any new technologies.

AI policy documents often read as though they were written in a vacuum. The DoD should begin any statement of AI principles by noting that the department is already bound to uphold the constitution, and abide by domestic and international law, and any use of AI should do the same. The task should then be to articulate principles that apply to the distinctive risks posed by AI, which cannot simply be derived from considering those overarching values. Those principles should take into account the distinctive nature of Defense, and should not simply amount to a rehashing of existing principles developed by other organisations. Finally, they should take into account the different constituencies that might be affected by Defense's actions, and in particular recognise that the extent to which these principles constrain Defense might depend on the circumstances and constituencies affected.

Discussion of Possible Principles

Before suggesting some principles that I think Defense should consider, I want to drive home the point about how the standard principles don't apply to defense in a straightforward way. I'll then go on to suggest some guiding principles that might pass that test.

Explainability

Most AI principle sets have some version of an 'explainability' principle—often linked to or differentiated from interpretability, transparency and so on. Rather than dig into the details of these different principles, I want to make some general points about how they apply to Defense.

Start with applications of AI that affect non-citizens, whether in a war-fighting or humanitarian context. Although international human rights law involves many proud statements of universal human rights, the reality is that international law as practised provides a meagre set of protections to people just in virtue of their humanity, and it is deeply implausible that either in the context of war or of humanitarian action, Defense would owe an explanation of how its algorithmic systems work to those affected by its decisions. It's not even plausible that there would be an obligation at international law to provide some overarching international body with insight into how Defense makes its decisions. Indeed, the US is regrettably ill-disposed towards international courts. And it is much more plausible that the standard of international law is one of *results* rather than processes. Different countries already vary so much in their decision-making processes—it's hard enough to gain any consensus on what results are unacceptable. Gaining consensus on the processes by which those results could be reached is surely impossible.

Might explainability still be an important principle for Defense operations that affect US citizens? Insofar as those citizens are members of the armed services, it's unclear why there would be a requirement to explain algorithmic decision-making to service-members when there is no requirement to explain any other kind of decision-making. That is, as long as the algorithmic system results in the delivery of an order that is not obviously unlawful, the standard expectation would be that orders are followed, without any explanation being owed. Of course, Defense has many civilian employees as well, and they are entitled to an explanation of certain kinds of decisions affecting them, in much the same way as any civilian employee of a government department would be—which is very likely already covered by existing employment law.

The most distinctive point at which an explanation of AI systems might be owed would be to the representatives of the citizens, in government. In order to ensure civilian control of the military, it is perhaps necessary to report to Congress in ways that make the operations of algorithms explicit. But it also seems plausible that there can be entirely successful civilian oversight of Defense without any AI systems deployed being explainable. How AI fits into the DoD's decision-making processes can clearly be regulated without explainability, as can the results of those processes.

Accountability and Contestability

Explainability, transparency, and interpretability are often separated from accountability and contestability. This is arguably a mistake—we care about explanation, transparency, and interpretability because they enable contestability, which is itself necessary for accountability. And you can't have accountability without all those preconditions. So it would be possible to just state that the central principle here is accountability.

Analytical nit-picking aside, however, it's clear that we can raise just the same questions for contestability as we did for explainability. Accountability however is not necessarily quite so variable dependent on the affected constituency. Even in war, the military must be accountable for its actions. Accountability is equally important, when fundamental rights are at stake, whether the rights are held by citizens or non-citizens (or at least, it's really important even in the latter case). However, this is an area in which the DoD is in a *better* position than private actors or even national governments. There already exist very clear structures of accountability within military organisations, which entail forms of group and command responsibility. Units are collectively responsible for the actions of their members, provided they act ‘intra vires’ (within the mutually-understood bounds of their collective endeavour). Commanders are responsible for the actions of their subordinates (subject to the same condition). Even when subordinates act ‘ultra vires’ commanders are still responsible to at least some degree. Where AI systems are used in a battle context, nothing changes. If anything, military applications are the ones where the thesis that AI systems dilute and diffuse responsibility is the least plausible, because the military is a highly structured collective acting according to a clear hierarchy and set of rules—they are the paradigmatic group agent—so the difference between having a human and an AI system at the ‘sharp end of the spear’ is relatively minimal.

Fairness and Privacy

Perhaps the two central concerns raised by AI systems have been that they might exacerbate discrimination by learning from data into which discriminatory practices are embedded, and that the data on which they are trained might be used without the consent of the originating parties. These principles seem particularly irrelevant when it comes to Defense operations that affect non-citizens. In humanitarian situations, the only good reason for DoD to take action that significantly impacts non-citizens is if there is an emergency that needs to be addressed, and the right to privacy must clearly give way if its doing so is necessary for lives to be saved. In war-fighting situations this is all the more clear. We could certainly argue that international human rights law protects civilians in war against certain kinds of invasions of their privacy, but such an argument is unlikely to gain much traction given the other much more serious depredations to which civilians are usually vulnerable in war.

Worries about fairness and discrimination in these contexts are vulnerable to the same arguments. Defense routinely makes discriminatory judgments in war-fighting situations—military-aged males, for example, are often on that basis alone considered to be legitimate targets when they appear in locations most frequently attended by other legitimate targets. ‘Profiling’ is a routine aspect of war-fighting, and is very plausibly permissible in light of its usefulness in mitigating risks. Of course, if the algorithmic allocation of humanitarian assistance had discriminatory effects, then that would be a potentially serious problem, but perhaps one that was overridable in the event that lives could thereby be saved. Still, clearly the DoD should strive to ensure, over time, that its use of AI systems in the allocation of humanitarian aid does not have discriminatory effects.

What then of the use of AI in applications that affect service-members? Again, members of the armed services enjoy weaker protections against invasions of their privacy by their service than would ordinary citizens. And some discrimination is clearly tolerated within the military—for example based on gender. Of course, Defense also has many civilian employees, and insofar as AI systems are used in ways that affect them, the same kinds of consideration would apply as apply to other government organisations when their actions affect their staff.

The net result of these arguments is that it would be very hard to articulate general principles governing explainability, accountability, fairness, and privacy, for all operations by the DoD. How they apply really depends very strictly on the affected constituency and the circumstances. One could write that DoD should ensure *to the extent feasible given the circumstances* that explainability, accountability, privacy, and fairness are satisfied. One could then give a detailed discussion of what is meant by this, somewhat along the lines of what is discussed here. One might reasonably question whether a principle qualified in this way is really a principle at all, however.

Alternatively, one could present the principles in this form: **DoD must ensure that adoption of AI systems does not prevent it from meeting existing obligations to render its decisions explainable to those affected by them, to be appropriately accountable for those decisions, and to make them in a non-discriminatory way, respecting individual privacy**. This would then presuppose an account of what those existing obligations are, so would again be of dubious value as a standalone principle.

Alternative Principles

Many of the risks posed by the deployment of AI are the same as the risks involved in any other operational decision taken by DoD. There is a general principle governing all morally risky conduct, which DoD must observe in its deployment of AI, the principle of necessity:

DoD must recognise the different degrees of moral risk in adopting AI, and proceed with adoption only when the benefits of doing so clearly justify the moral risks, and where no other less risky alternative is available to realise comparable benefits.

Note that this principle applies both to the choice to use AI in the first place, and to the way in which the AI system is used. The first point is important: there is a headlong rush to adopt AI systems, but sometimes it simply is not worth the risk (the adoption of facial recognition systems now seems to be such a case). The second point matters too: any automated systems must be subjected to continual audit and review (as should any DoD systems).

Another principle that is more specific to AI (though would really apply, mutatis mutandis, to any new technology or process), would focus our attention on the risks that adoption of AI systems generate for human behaviour. It is well-documented that implementing automated systems risks leading to de-skilling and automation bias (where human operators defer to automated systems). This should be of particular concern to DoD, since it depends so heavily on the skilled judgment of its members. There should therefore be a principle along these lines:

DoD must ensure that its representatives are trained to develop and exercise their own critical judgment when operating in partnership with AI systems.

Killer Robots?

I haven't yet mentioned lethal autonomous weapons directly. It's worth briefly pausing to ask whether claims made by organisations like the Campaign to Stop Killer Robots are plausible. One candidate principle that might be proposed here is the following:

DoD must ensure that decisions over life and death are never taken by an AI system.

This principle would caution against allowing AI systems to take action that is expected to result in death without an intervening decision being made by a human operator. It would likely not apply in clearly low-risk cases where human intervention renders defence infeasible—for example Phalanx missiles.

There may be pragmatic reasons to endorse a principle such as this. In particular, it may help DoD gain public trust in its adoption of AI systems, since there seems to be considerable public support for this principle. It might also enable international coordination around a ban on autonomous weapons, which might help to limit proliferation, and constrain the moral costs of war. And of course existing and near-future AI technology is plausibly not able to abide by the laws of war—so this principle would just be a special case of the principle of necessity above.

Are there non-instrumental reasons to think that autonomous weapons are especially objectionable? If there are, I think they would have to be grounded in the considerations adduced above—to do with accountability, acting for the right reasons, care and attention, or avoiding discrimination. But I've already argued that military organisations have robust structures of accountability in place, and in any event are best viewed as corporate agents. AI systems do not change that. I am sceptical about whether we should really care about the 'acting for the right reasons' point in a military context. After all, we make no effort to explore the reasons on which individual combatants are acting, and in fact we typically encourage them to act on reasons that might not be particularly relevant to the overall justification for their actions—namely, many combatants are motivated by a desire to protect or avenge their friends, which while morally important in its own right, is strictly not relevant to the justice of the cause for which the war is fought, and so on. The idea that human combatants kill with care and attention for their victims is unrealistic at best; no doubt many do manage to maintain such attitudes at least some of the time, in some conflicts. But it is surely understandable that many people find their compassion has limits in the circumstances of war. And as already argued, in the radical uncertainty faced by combatants in war, reliance on heuristics is inevitable; AI systems would quite plausibly be less discriminatory than humans in the same situation, because they would at least be able to draw on much more information when selecting targets.

In my view, then, realistic AI systems would only ever be tools used by the military unit to which they are assigned. They should be assessed in that light. There is nothing intrinsically wrong with them, but they should not be adopted unless they satisfy the principle of necessity described above. And there may be pragmatic grounds to reject them even if they did satisfy that principle.

However, I think that DoD is governed by a further, positive principle (in some respects the flipside of the principle of necessity):

DoD must adopt AI systems where doing so would reduce the risk of civilian casualties.

International law requires militaries to take feasible precautions in attack. If there are technological precautions that could reduce the risk of civilian casualties, then they must be taken—they are not optional. So if DoD is able to develop weapons that are able to disarm themselves on realising that the risk of civilian casualties has significantly increased in the seconds since their launch, then it has very good reason to do so. Likewise if technologies can be developed that better enable identifying where civilians are in the first place.

AI and International Security

I have least to say on this point; other scholars versed in security studies and international relations will be better placed to advise. But it’s clearly crucial for the DoD to pay close attention to the distinctive security threats that AI makes possible, and to consider whether distinct principles are necessary to address those threats. However, in this respect AI really is just part of a suite of cyber security risks that Defense should address together.