# AFF ANSWERS

## AT Link

### Governance/Regulation

#### Regulation is key to capture the benefits of AI – solves laundry list of societal and security issues

Tzimas ’21 [Themistoklis; 2021; Faculty of Law at the Aristotle University of Thessaloniki; Legal and Ethical Challenges of Artificial Intelligence from an International Law Perspective, “Chapter 2: The Expectations and Risks from AI,” p. 9-32]

Therefore, it is only natural to be at least skeptical towards a future with entities possessing equal or superior intelligence and levels of autonomy; the prospect even of existential risk looms as possible.7 AI that will have reached or surpassed our level of intelligence make us wonder why would highly autonomous and intelligent AI want to give up control back to its original creators?8 Why remain contained in pre-deﬁned goals set for it by us, humans?Even AI in its current form and narrow intelligence poses risks because of its embedded-ness in an ever-growing number of crucial aspects of our lives. The role of AI in military, ﬁnancial,9 health, educational, environmental, governance networks-among others—are areas where risk generated by AI—even limited— autonomy can be diffused through non-linear networks, with signiﬁcant impact— even systemic.10 The answer therefore to the question whether AI brings risk with it is yes; as Eliezer Yudkowski comments the greatest of them all is that people conclude too early that they understand it11 or that they assume that they can achieve it without necessarily having acquired complete and thorough understanding of what intelli- gence means.12 Our projection of our—lack of complete—understanding of the concept of intelligence on AI is owed to our lack of complete comprehension of human intelligence too, which is partially covered by the prevalent and until now self- obvious, anthropomorphism because of which we tend to identify higher intelligence with the human mind.Yudkowski again however suggests that AI “refers to a vastly greater space of possibilities than does the term “Homo sapiens.” When we talk about “AIs” we are really talking about minds-in-general, or optimization processes in general. Imagine a map of mind design space. In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the space of minds-in-general. The entire map ﬂoats in a still vaster space, the space of optimization processes.”13 Regardless of what our well-established ideas are, there are many, different intelligences and even more signiﬁcantly, there are potentially, different intelli- gences equally or even more evolved than human.From such a perspective, the unprecedented—ness of potential AI developments and the mystery surrounding them emerges as not only the outcome of pop culture but of a radical transformation of our—until recently—self—obvious identiﬁcation of humanity with highly evolved and dominant intelligence.14 The lack of understanding of intelligence and therefore of AI may be frightening but does not lead necessarily to regulation—at least to a proper one. We could even be led into making potentially catastrophic choices, on the basis of false assumptions. On top of our lack of understanding, we should add a sentiment of anxiety as well as of expectations, which intensiﬁes as an atmosphere of emergency and of expected groundbreaking developments grows. The most graphic description of this feeling is the potential of a moment of singularity, as mentioned above according to the description by Vinge and Kurzweil. As the mathematician I. J. Good–Alan Turing’s colleague in the team of the latter during World War II—has put it: “Let an ultraintelligent machine be deﬁned as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the ﬁrst ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”15 This is in a nutshell the moment of singularity. The estimates currently foresee the emergence of ultra or super intelligence—as it is currently labelled—or in other words of singularity, somewhere between 20 and 50 years from today, further raising the sentiment of emergency.16 We cannot even foretell with precision how singularity would look like but we know that because of its expected groundbreaking impact, both states and private entities compete towards gaining the upper hand in the prospect of the singularity.17 Despite the fact that such predictions have been proven rather optimistic in the past18 and therefore up to some extent inaccurate, there are reasons to assume that their materialization will take place and that the urgency of regulation will be proven realistic. After all, part of the disappointments from AI should be blamed on the fact that certain activities and standards, which were considered as epitomes of human intelligence have been surpassed by AI, only to indicate that they were not eventu- ally satisfactory thresholds for the surpassing of human intelligence.19 Partially because of AI progress we realize that human intelligence and its thresholds are much more complicated than assumed in the past. The vastness’s of deﬁnitions of intelligence, as well as its etymological roots are enlightening of the difﬁculties: “to gather, to collect, to assemble or to choose, and to form an impression, thus leading one to ﬁnally understand, perceive, or know”.20 As with other relevant concepts, the truth is that until recently our main way to approach intelligence for far too long was “we know it, when we see it”. AI is an additional reason for looking deeper into intelligence and the more we examine it, the most complicated it seems. The combination of lack of complete understanding of intelligence, the unpredictability of AI, its rapid evolution and the prospect of singularity explain both the fascination and the fear from AI. Once the latter emerges, we have no real knowledge about what will happen next but only speculations, which until recently belonged to the area of science ﬁction. We are for example pretty conﬁdent that the speed of AI intelligence growth will accelerate, once self—improvement will have been achieved. The expected or possible chain of events will begin from AI capacity to re-write its own algorithms and exponentially self—improve, surpassing human intelligence, which lacks the capacity of such rapid self—improvement and setting its own goals.21 We can somehow guess the speed of AGI and ASI evolution and possibly some of its initial steps but we cannot guess the directions that such AI will choose to follow and the characteristics that it will demonstrate. Practically, we credibly guess the prospects of AI beyond a certain level of development. Two existential issues could emerge: ﬁrst, an imbalance of intelligence at our expense—with us, humans becoming the inferior species—in favor of non-biological entities and secondly a lack of even fundamental conceptual communication between the two most intelligent “species”. Both of them heighten the fear of irreversible changes, once we lose the possession of the superior intelligence.22 However, we need to consider the expectations as well. The positive side focuses on the so-called friendly AI, meaning AI which will beneﬁt and not harm humans, thanks to its advanced intelligence.23 AI bears the promise of signiﬁcantly enhancing human life on various aspects, beginning from the already existing, narrow applications. The enhanced automation24 in the industry and the shift to autonomy,25 the take—over by AI of tasks even at the service sector which can be considered as “tedious”—i.e. in the banking sector—climate and weather forecasting, disaster response,26 the potentially better cooperation among different actors in complicated matters such as in matters of information, geopolitics and international relations, logistics, resources ex.27 The realization of the positive expectations depends up to some extent upon the complementarity or not, of AI with human intelligence. However, what friendly AI will bring in our societies constitutes a matter of debate, given our lack of unanimous approach on what should be considered as beneﬁcial and therefore friendly to humans—as is analyzed in the next chapter.Friendly AI for example bears the prospect of freeing us from hard labor or even further from unwanted labor; of generating further economic growth; of dealing in unbiased, speedy, effective and cheaper ways with sectors such as policing, justice, health, environmental crisis, natural disasters, education, governance, defense and several more of them which necessitate decision-making, with the involvement of sophisticated intelligence.The synergies between human intelligence and AI “promise” the enhancement of humans in most of their aspects. Such synergies may remain external—humans using AI as external to themselves, in terms of analysis, forecasts, decision—making and in general as a type of assistant-28 or may evolve into the merging of the two forms of intelligence either temporarily or permanently.The second profoundly enters humanity, existentially—speaking, into uncharted waters. Elon Musk argues in favor of “having some sort of merger of biological intelligence and machine intelligence” and his company “Neuralink” aims at implanting chips in human brain. Musk argues that through this way humans will keep artiﬁcial intelligence under control.29 The proposition is that of “mind design”, with humans playing the role that God had according to theologies.30 While the temptation is strong—exceeding human mind’s capacities, far beyond what nature “created”, by acquiring the capacity for example to connect directly to the cyberspace or to break the barriers of biology31—the risks are signiﬁcant too: what if a microchip malfunction? Will such a brain be usurped or become captive to malfunctioning AI?The merging of the two intelligences is most likely to evolve initially by invoking medical reasons, instead of human enhancement. But the merging of the two will most likely continue, as after all the limits between healing and enhancement are most often blurry. This development will give rise, as is analyzed below, to signif- icant questions and issues, the most of crucial of which is the setting of a threshold for the prevalence of the human aspect of intelligence over the artiﬁcial one.Human nature is historically improved, enhanced, healed and now, potentially even re-designed in the future.32 Can a “medical science” endorsing such a goal be ethically acceptable and if yes, under what conditions, when, for whom and by what means? The answers are more difﬁcult than it seems. As the World Health Organi- zation—WHO—provides in its constitution, “Health is a state of complete physical, mental and social well-being and not merely the absence of disease or inﬁrmity”.33 Therefore, why discourage science which aims at human-enhancement, even reaching the levels of post-humanism?34 Or if restrictions are to be imposed on human enhancement, on what ethics and laws will they be justiﬁed? How ethically acceptable is it to prohibit or delay technological evolution, which among several other magniﬁcent achievements, promises to treat death as a disease and cure it, by reducing soul to self, self to mind, and mind to brain, which will then be preserved as a “softwarized” program in a hardware other than the human body?35 After all, “According to the strong artiﬁcial intelligence program there is no fundamental difference between computers and brains: a computer is different machinery than a person in terms of speed and memory capacity.”36 While such a scientiﬁc development and the ones leading potentially to it will be undoubtedly, groundbreaking technologically-speaking, is it actually—ethically- speaking—as ambivalent as it may sound or is it already justiﬁed by our well— rooted human-centrism?37 Secular humanism may have very well outdated religious beliefs about afterlife in the area of science but has not diminished the hope for immortality; on the contrary, science, implicitly or explicitly predicts that matter can in various ways surpass death, albeit by means which belong in the realm of scientiﬁc proof, instead of that of metaphysical belief.38 If this is the philosophical case, the quest for immortality becomes ethically acceptable; it can be considered as embedded both in the existential anxiety of humans, as well as in the human-centrism of secular philosophical and political victory over the dei-centric approach to the world and to our existence. From another perspective of course and for the not that distant philosophical reasons, the quest for immortality becomes ethically ambiguous or even unacceptable.39 By seeking endless life we may miss all these that make life worth living in the framework of ﬁniteness. As the gerontologist Paul Hayﬂick cautioned “Given the possibility that you could replace all your parts, including your brain, then you lose your self-identity, your self-recognition. You lose who you are! You are who you are because of your memory.”40 In other words, once we begin to integrate the two types of intelligence, within ourselves, until when and how we will be sure that it is human intelligence that guides us, instead of the AI? And if we are not guided completely or—even further—at all by human intelligence but on the contrary we are guided by AI which we have embodied and which is trained by our human intelligence, will we be remaining humans or we will have evolved to some type of meta-human or transhumant species, being different persons as well?41 AI promises tor threatens to offer a solution by breaking down our consciousness into small “particles” of information—simplistically speaking—which can then be “software-ized” and therefore “uploaded” into different forms of physical or non-physical existence. Diane Ackerman states that “The brain is silent, the brain is dark, the brain tastes nothing, the brain hears nothing. All it receives are electrical impulses--not the sumptuous chocolate melting sweetly, not the oboe solo like the ﬂight of a bird, not the pastel pink and lavender sunset over the coral reef--only impulses.”42 Therefore, all that is needed—although it is of course much more complicated than we can imagine—is a way to code and reproduce such impulses. Even if we consider that without death, we will no more be humans but something else, why should we remain humans once technologies allow us be something “more”, in the sense of an enhanced version of “being”? Why are we to remain bound by biological evolution if we can re-design it and our future form of existence? Why not try to achieve the major breakthrough, the anticipated or hoped digita- lization of the human mind, which promises immortality of consciousness via the cyberspace or artiﬁcial bodies: the uploading of our consciousness so that it can live on forever, turning death into an optional condition.43 Either through an artiﬁcial body or emulation-a living, conscious avatar—we hope—or fear—that the domain of immortality will be within reach. It is the prospect of a “substrate-independent minds,” in which human and machine consciousness will merge, transcending biological limits of time, space and mem- ory” that fascinates us.44 As Anders Sandberg explained “The point of brain emulation is to recreate the function of the original brain: if ‘run’ it will be able to think and act as the original,” he says. Progress has been slow but steady. “We are now able to take small brain tissue samples and map them in 3D. These are at exquisite resolution, but the blocks are just a few microns across. We can run simulations of the size of a mouse brain on supercomputers—but we do not have the total connectivity yet. As methods improve, I expect to see automatic conversion of scanned tissue into models that can be run. The different parts exist, but so far there is no pipeline from brains to emulations.”45 The emulation is different from a simulation in the sense that the former mimics not only the outward outcome but also the “internal causal dynamics”, so that the emulated system and in this particular case the human mind behaves as the original.46 Obviously, this is a challenging task: we need to understand the human brain with the help of computational neuroscience and combine simpliﬁed parts such as simulated neurons with network structures so that the patterns of the brain are comprehended. We must combine effectively “biological realism (attempting to be faithful to biology), completeness (using all available empirical data about the system), tractability (the possibility of quantitative or qualitative simulation) and understanding (producing a compressed representation of the salient aspects of the system in the mind of the experimenter)”.47 The technological challenges are vast. Technologically speaking, the whole concept is based on some assumptions which must be proven both accurate and feasible.48 We must achieve technology capable of scanning completely the human brain, of creating software on the basis of the acquired information from its scanning and of the interpretation of information and the hardware which will be capable of uploading or downloading such software.49 The steps within these procedures are equally challenging. Their detailed analysis evades the scope of this book. Some critical questions—they are further analyzed in the next chapters—emerge however: how will we interpret free will in emulation? What will be the impact of the environment and of what environment? How will be missing parts of the human brain re-constructed and emulated? What will be the status of the several emulations which will be created—i.e. failed attempts or emulations of parts of the human brain—in the course of the search for a complete and functioning emulation? Will they be considered as “persons” and therefore as having some right or will they be considered as mere objects in an experimental lab? How are we going to decode the actual subjective sentiments of these emulations? Essentially, are emulations the humans “themselves” who are emulated or a different person? Even further what will human and person mean in the era of emulation? From a different perspective, the victory over death may be seen as a danger of mass extinction, absorption or de-humanization. In this new, vast universe of emulations will there be place for humans?50 From the above—mentioned discussion, it becomes obvious that at a large extent, the prospect of risk or of expectation is a matter of perspective, for which there is no unanimous agreement in the present. This may be the greatest danger of all, for which Asimov warned us: unleashing technology while we cannot communicate among us, in the face of it. The existential prospect as well as the risks by AI may self-evidently emerge from technological advances but are determined on the basis of politico—philosophical or in the wider sense, ethical assumptions. This is where the need for legal regulation steps in. Such a need was often underestimated in the past in favor of a solely technologically oriented approach—although exceptions raising issues other than technological can be found too.51 The gradual raising of ethic—political, philosoph- ical and legal issues constitutes a rather recent development, partially because of the realization of the proximity of the risks and of the expectations. The public debate is often divided between two “contradictory” views: fear of AI or enthusiastic optimism. The opinions of the experts differ respectively. Kurzweil, who has come with a prediction for a date for the emergence of singularity—until 2045—expects such a development in a positive way: “What’s actually happening is [machines] are powering all of us,” Kurzweil said during the SXSW interview. “They’re making us smarter. They may not yet be inside our bodies, but, by the 2030s, we will connect our neocortex, the part of our brain where we do our thinking, to the cloud.”52 In a well-known article—issued on the occasion of a ﬁlm—Stephen Hawking, Max Tegmark, Stuart Russell, and Frank Wilczek shared a moderate position: “The potential beneﬁts are huge; everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magniﬁed by the tools AI may provide, but the eradication of war, disease, and poverty would be high on anyone’s list. Success in creating AI would be the biggest event in human history. . . Unfortunately, it might also be the last, unless we learn how to avoid the risks.”53

### AWS = Moral

#### Autonomous weapons better for ethics- reduces unnecessary killing

Galliott, Jai. "Humans, autonomous systems, and killing in war." Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics. IGI Global, 2021. 240-257. [AJL]

However, not all are convinced by the argument that autonomous systems present a moral problem in making it easier to indiscriminately and disproportionately kill, even at the higher end of the spectrum of autonomy. Daniel Brunstetter and Megan Braun (2011, p. 339) argue that semi-autonomous robotic systems are subject to the same jus in bello requirements as other weapons used in war, but that their +XPDQV$XWRQRPRXV6\VWHPVDQG.LOOLQJLQ:DU technological advantages coupled with the removal of risk to soldiers means that they should, at the least in theory, make satisfying the principles of discrimination and proportionality an easier task and perhaps make operators more reluctant to kill in situations where doubt exists as to the legitimacy of the potential victim of aggression. They say that the in the case of surveillance, at the very least, the distance or what they call ‘separation factor’, arguably offers an increased level of control over lethal targeting decisions and ought to actually reduce the emotional toll and unnecessary killing (Brunstetter and Braun 2011, p. 339). They regard a drone operator’s ability to confer with a superior officer as being a critical factor encouraging ethical decision making in war. In some instances, this may be the case, and may even apply in the case of highly autonomous systems if the relevant coding and engineering is sufficiently detailed and comprehensive to account for all relevant morally relevant inputs and outputs, but in others, it might be that having a mission commander or a test overseeing the operator’s/programmer’s actions only places additional pressure on them to perpetrate lethal acts, just as the sergeants walking the trenches of WWI aimed to encourage reluctant soldiers to kill. Yet Christian Enemark (2013) also questions some of the assumptions relied on here. He says that there is reason to suppose that being physically absent from the battlefield is more conducive to discrimination (Enemark 2013). In his view, the removal of risk allows decisions to be made in a more deliberate manner and also removes anger and emotion that he thinks might otherwise lead to morally unsanctioned killings. That is, if a drone operator working from a desk in Nevada encounters the enemy, adherence to jus in bello protocol should improve as the operator is at little or no personal risk. It could be suggested, however, that if an operator or technical contributors is so emotionally removed, they are in fact likely to develop the sort disengagement referred to above or an even more morally concerning callousness. In case of highly autonomous system with little input other than through code, the concern that it such a callousness might pervade said code and perhaps even go unnoticed by virtue of being concealed within the system and being evident only from its highly complex actions, having even more severe and long-lasting consequences.

#### Autonomous weapons are at best more ethical than humans and at worse just as ethical as humans

Gunkel 17 (David, Professor of Communication Studies at Northern Illinois University. 2017 “Mind the gap: responsible robotics and the problem of responsibility” Ethics and Information Technology. doi:10.1007/s10676-017-9428-2)-qcl

Conversely, we can entertain the possibility of what has been called “machine ethics” just as we had previously done for other non-human entities, like animals (Singer 1975). And there has, in fact, been a number of recent proposals addressing this opportunity. Wallach and Allen (2009, p. 4), for example, not only predict that “there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight” but use this fact as justification for developing “moral machines,” advanced technological systems that are able to respond to morally challenging situations. Anderson and Anderson (2011) take things one step further. They not only identify a pressing need to consider the moral responsibilities and capabilities of increasingly autonomous systems but have even suggested that “computers might be better at following an ethical theory than most humans,” because humans “tend to be inconsistent in their reasoning” and “have difficulty juggling the complexities of ethical decision-making” owing to the sheer volume of data that need to be taken into account and processed (Anderson and Anderson 2007, p. 5). These proposals, it is important to point out, do not necessarily require that we first resolve the “big questions” of AGI (Artificial General Intelligence), robot sentience, or machine consciousness. As Wallach (2015, p. 242) points out, these kinds of machines need only be “functionally moral.” That is, they can be designed to be “capable of making ethical determinations…even if they have little or no actual understanding of the tasks they perform.” The precedent for this way of thinking can be found in corporate law and business ethics. Corporations are, according to both national and international law, legal persons (French 1979). They are considered “persons” (which is, we should recall, a moral classification and not an ontological category) not because they are conscious entities like we assume ourselves to be, but because social circumstances make it necessary to assign personhood to these artificial entities for the purposes of social organization and jurisprudence. Consequently, if entirely artificial and human fabricated entities, like Google or IBM, are legal persons with associated social responsibilities, it would be possible, it seems, to extend the same moral and legal considerations to an AI or robot like Google’s DeepMind or IBM’s Watson. The question, it is important to point out, is not whether these mechanisms are or could be “natural persons” with what is assumed to be “genuine” moral status; the question is whether it would make sense and be expedient, from both a legal and moral perspective, to treat these mechanisms as persons in the same way that we currently do for corporations, organizations and other human artifacts. Once again, this decision sounds reasonable and justified. It extends both moral and legal responsibility to these other socially aware and interactive entities and recognizes, following the predictions of Wiener (1988, p. 16), that the social situation of the future will involve not just human-tohuman interactions but relationships between humans and machines and machines and machines. But this shift in perspective also has significant costs. First, it requires that we rethink everything we thought we knew about ourselves, technology, and ethics. It entails that we learn to think beyond human exceptionalism, technological instrumentalism, and many of the other -isms that have helped us make sense of our world and our place in it. In effect, it calls for a thorough reconceptualization of who or what should be considered a legitimate center of moral concern and why Second, robots that are designed to follow rules and operate within the boundaries of some kind of programmed restraint, might turn out to be something other than what is typically recognized as a responsible agent. Winograd (1990, pp. 182–183), for example, warns against something he calls “the bureaucracy of mind,” “where rules can be followed without interpretive judgments.” “When a person,” Winograd (1990, p. 183) argues, “views his or her job as the correct application of a set of rules (whether human-invoked or computerbased), there is a loss of personal responsibility or commitment. The ‘I just follow the rules’ of the bureaucratic clerk has its direct analog in ‘That’s what the knowledge base says.’ The individual is not committed to appropriate results, but to faithful application of procedures.” Coeckelbergh (2010, p. 236) paints a potentially more disturbing picture. For him, the problem is not the advent of “artificial bureaucrats” but “psychopathic robots.” The term “psychopathy” has traditionally been used to name a kind of personality disorder characterized by an abnormal lack of empathy which is masked by an ability to appear normal in most social situations. The functional morality, like that specified by Anderson and Anderson and Wallach and Allen, intentionally designs and produces what are arguably “artificial psychopaths”—robots that have no capacity for empathy but which follow rules and in doing so can appear to behave in morally appropriate ways. These psychopathic machines would, Coeckelbergh (2010, p. 236) argues, “follow rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents—i.e. agents that follow rules without being moved by moral concerns—and they would even lack the capacity to discern what is of value. They would be morally blind.”4 Efforts in “machine ethics” (or whatever other nomenclature comes to be utilized to name this development) effectively seek to widen the circle of moral subjects to include what had been previously excluded and marginalized as mere neutral instruments of human action. This is, it is important to note, not some blanket statement that would turn everything that was a tool into a moral subject. It is the recognition, following Marx, that not everything technological is reducible to a tool and that some devices—what Marx called “machines” and what Winner calls “autonomous technology”—might need to be programmed in such a way as to behave reasonably and responsibly for the sake of respecting human individuals and communities. This proposal has the obvious advantage of responding to moral intuitions: if it is the machine that is making the decision and taking action in the world with little or no direct human oversight, it would only make sense to hold it accountable (or at least partially accountable) for the actions it deploys and to design it with some form of constraint in order to control for possible bad outcomes. But doing so has considerable costs. Even if we bracket the questions of AGI, super intelligence, and machine consciousness; designing robotic systems that follow prescribed rules might provide the right kind of external behaviors but the motivations for doing so might be lacking. “Even if,” Sharkey (2012, p. 121) writes in a consideration of autonomous weapons, “a robot was fully equipped with all the rules from the Laws of War, and had, by some mysterious means, a way of making the same discriminations as humans make, it could not be ethical in the same way as is an ethical human. Ask any judge what they think about blindly following rules and laws.” Consequently, what we actually get from these efforts might be something very different from (and maybe even worse than) what we had hoped to achieve.

#### Autonomous weapons are moral – the ability for targets to fight back against autonomous weapons mean they respect the target’s autonomy

**Young 21** (Garry director at the GW Institute of Public Policy, 03-29-2021, accessed on 6-21-2022, Ethics and Information Technology, “On the indignity of killer robots. Ethics and Information Technology,” 23(3), 473–482. https://doi.org/10.1007/s10676-021-09590-2 , pp. 6

A strong rebuttal of the indignity argument denies the truth of P3 (that the deployment of killer robots disrespects the dignity of combatants). Positioning ourselves once more behind a veil of ignorance, we again ask: what would military commanders be agreeing to if they were to agree to the permissibility of killer robots? We know that they would be agreeing to deploy autonomous weapons whose decision making cannot be constrained by recognition respect. In deciding whether combatants live or die, the killer robots would be ‘treating’ them as objects and not as moral agents with inherent sortal dignity. In other words, they would be processing combatants in a manner that would be no diferent to any other object in their environment. Given this, the charge is that we (qua military commanders), by agreeing to deploy killer robots in this way, would be treating combatants in a manner that disrespect their sortal dignity. It is this claim that the strong rebuttal challenges. Permitting killer robots does not deny combatants the opportunity to fght back against these automated weapons, and therefore act as moral agents. The weapons themselves may not be capable of respecting the inherent dignity of the combatant they target, making their deaths appear arbitrary (Amoroso, 2017) but, in deciding to deploy such weapons, we are capable of recognizing the inherent dignity of the combatants these weapons will eventually target. After all, even if we accept that sortal dignity is inherent, what counts as an afront to this dignity is not immuable. Instead, it is constructed, and forms part of what Killmister (2017) refers to as social indignity. Therefore, from behind the veil of ignorance, we (qua the community of military commanders) could agree (socially construct and endorse the view) that killer robots are not an afront to sortal dignity because their deployment does not prevent combatants from acting as moral agents (exercising their rational autonomy) and quite possible neutralizing the killer robots in return: a fact we recognize and respect.11 Consequently, P3 of the indignity argument is false, meaning that C(ii) does not necessarily follow. Given the stronger rebuttal of the indignity argument, if P3 is rejected then C(iii)—the claim that the death of a combatant, as a consequence of P3, amounts to an undignifed death—is likewise rejected. In the case of the weaker rebuttal, however, where P3 is not rejected, might a case still be made for the truth of C(iii): that even when treating a lack of respect for the dignity of combatants as a pro tanto wrong, the afront to the combatant’s dignity nevertheless results in an undignifed death? I do not believe so, as I intend to show in the next section. By drawing on two examples from fction I defend the claim that one can preserve one’s outward dignity in the face of indignity but, also, that the preservation of dignity supports the stronger rebuttal’s claim that recognition respect would be bestowed on combatants faced with an assault from killer robots from the community of military commanders (as well as others), adding weight to the claim that the deployment of killer robots is not in fact undignifed. Either way, C(iii) is undermined.

### AI is fixable/ethical

#### AI could become more ethical than humans and provide more time for humans to benefit society

Bossmann 16 Julia Bossmann, 10-21-2016, "Top 9 ethical issues in artificial intelligence," World Economic Forum, <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> [AJL]

Optimizing logistics, detecting fraud, composing art, conducting research, providing translations: intelligent machine systems are transforming our lives for the better. As these systems become more capable, our world becomes more efficient and consequently richer. Tech giants such as Alphabet, Amazon, Facebook, IBM and Microsoft – as well as individuals like Stephen Hawking and Elon Musk – believe that now is the right time to talk about the nearly boundless landscape of artificial intelligence. In many ways, this is just as much a new frontier for ethics and risk assessment as it is for emerging technology. So which issues and conversations keep AI experts up at night? 1. Unemployment. What happens after the end of jobs? The hierarchy of labour is concerned primarily with automation. As we’ve invented ways to automate jobs, we could create room for people to assume more complex roles, moving from the physical work that dominated the pre-industrial globe to the cognitive labour that characterizes strategic and administrative work in our globalized society. Look at trucking: it currently employs millions of individuals in the United States alone. What will happen to them if the self-driving trucks promised by Tesla’s Elon Musk become widely available in the next decade? But on the other hand, if we consider the lower risk of accidents, self-driving trucks seem like an ethical choice. The same scenario could happen to office workers, as well as to the majority of the workforce in developed countries. Have you read? Artificial Intelligence Collides with Patent Law Robot inventors are on the rise. But are they welcomed by the patent system? Artificial intelligence could be our saviour, according to the CEO of Google This is where we come to the question of how we are going to spend our time. Most people still rely on selling their time to have enough income to sustain themselves and their families. We can only hope that this opportunity will enable people to find meaning in non-labour activities, such as caring for their families, engaging with their communities and learning new ways to contribute to human society. If we succeed with the transition, one day we might look back and think that it was barbaric that human beings were required to sell the majority of their waking time just to be able to live. 2. Inequality. How do we distribute the wealth created by machines? Our economic system is based on compensation for contribution to the economy, often assessed using an hourly wage. The majority of companies are still dependent on hourly work when it comes to products and services. But by using artificial intelligence, a company can drastically cut down on relying on the human workforce, and this means that revenues will go to fewer people. Consequently, individuals who have ownership in AI-driven companies will make all the money. We are already seeing a widening wealth gap, where start-up founders take home a large portion of the economic surplus they create. In 2014, roughly the same revenues were generated by the three biggest companies in Detroit and the three biggest companies in Silicon Valley ... only in Silicon Valley there were 10 times fewer employees. If we’re truly imagining a post-work society, how do we structure a fair post-labour economy? 3. Humanity. How do machines affect our behaviour and interaction? Artificially intelligent bots are becoming better and better at modelling human conversation and relationships. In 2015, a bot named Eugene Goostman won the Turing Challenge for the first time. In this challenge, human raters used text input to chat with an unknown entity, then guessed whether they had been chatting with a human or a machine. Eugene Goostman fooled more than half of the human raters into thinking they had been talking to a human being. This milestone is only the start of an age where we will frequently interact with machines as if they are humans; whether in customer service or sales. While humans are limited in the attention and kindness that they can expend on another person, artificial bots can channel virtually unlimited resources into building relationships. Even though not many of us are aware of this, we are already witnesses to how machines can trigger the reward centres in the human brain. Just look at click-bait headlines and video games. These headlines are often optimized with A/B testing, a rudimentary form of algorithmic optimization for content to capture our attention. This and other methods are used to make numerous video and mobile games become addictive. Tech addiction is the new frontier of human dependency. On the other hand, maybe we can think of a different use for software, which has already become effective at directing human attention and triggering certain actions. When used right, this could evolve into an opportunity to nudge society towards more beneficial behavior. However, in the wrong hands it could prove detrimental. 4. Artificial stupidity. How can we guard against mistakes? Intelligence comes from learning, whether you’re human or machine. Systems usually have a training phase in which they "learn" to detect the right patterns and act according to their input. Once a system is fully trained, it can then go into test phase, where it is hit with more examples and we see how it performs. Obviously, the training phase cannot cover all possible examples that a system may deal with in the real world. These systems can be fooled in ways that humans wouldn't be. For example, random dot patterns can lead a machine to “see” things that aren’t there. If we rely on AI to bring us into a new world of labour, security and efficiency, we need to ensure that the machine performs as planned, and that people can’t overpower it to use it for their own ends.

AI Researchers are beginning to recognize biases and resolve them---means the squo will solve the links

Berreby 20 [David Berreby, 11-22-2020, accessed on 6-25-2022, The New York Times, "Can We Make Our Robots Less Biased Than We Are?", <https://www.nytimes.com/2020/11/22/science/artificial-intelligence-robots-racism-police.html>] -os-

On a summer night in Dallas in 2016, a bomb-handling robot made technological history. Police officers had attached roughly a pound of C-4 explosive to it, steered the device up to a wall near an active shooter and detonated the charge. In the explosion, the assailant, Micah Xavier Johnson, became the first person in the United States to be killed by a police robot. Afterward, then-Dallas Police Chief David Brown called the decision sound. Before the robot attacked, Mr. Johnson had shot five officers dead, wounded nine others and hit two civilians, and negotiations had stalled. Sending the machine was safer than sending in human officers, Mr. Brown said. But some robotics researchers were troubled. “Bomb squad” robots are marketed as tools for safely disposing of bombs, not for delivering them to targets. (In 2018, police officers in Dixmont, Maine, ended a shootout in a similar manner.). Their profession had supplied the police with a new form of lethal weapon, and in its first use as such, it had killed a Black man. “A key facet of the case is the man happened to be African-American,” Ayanna Howard, a robotics researcher at Georgia Tech, and Jason Borenstein, a colleague in the university’s school of public policy, wrote in a 2017 paper titled “The Ugly Truth About Ourselves and Our Robot Creations” in the journal Science and Engineering Ethics. Like almost all police robots in use today, the Dallas device was a straightforward remote-control platform. But more sophisticated robots are being developed in labs around the world, and they will use artificial intelligence to do much more. A robot with algorithms for, say, facial recognition, or predicting people’s actions, or deciding on its own to fire “nonlethal” projectiles is a robot that many researchers find problematic. The reason: Many of today’s algorithms are biased against people of color and others who are unlike the white, male, affluent and able-bodied designers of most computer and robot systems. While Mr. Johnson’s death resulted from a human decision, in the future such a decision might be made by a robot — one created by humans, with their flaws in judgment baked in. “Given the current tensions arising from police shootings of African-American men from Ferguson to Baton Rouge,” Dr. Howard, a leader of the organization Black in Robotics, and Dr. Borenstein wrote, “it is disconcerting that robot peacekeepers, including police and military robots, will, at some point, be given increased freedom to decide whether to take a human life, especially if problems related to bias have not been resolved.” Last summer, hundreds of A.I. and robotics researchers signed statements committing themselves to changing the way their fields work. One statement, from the organization Black in Computing, sounded an alarm that “the technologies we help create to benefit society are also disrupting Black communities through the proliferation of racial profiling.” Another manifesto, “No Justice, No Robots,” commits its signers to refusing to work with or for law enforcement agencies. Over the past decade, evidence has accumulated that “bias is the original sin of A.I,” Dr. Howard notes in her 2020 audiobook, “Sex, Race and Robots.” Facial-recognition systems have been shown to be more accurate in identifying white faces than those of other people. (In January, one such system told the Detroit police that it had matched photos of a suspected thief with the driver’s license photo of Robert Julian-Borchak Williams, a Black man with no connection to the crime.) There are A.I. systems enabling self-driving cars to detect pedestrians — last year Benjamin Wilson of Georgia Tech and his colleagues found that eight such systems were worse at recognizing people with darker skin tones than paler ones. Joy Buolamwini, the founder of the Algorithmic Justice League and a graduate researcher at the M.I.T. Media Lab, has encountered interactive robots at two different laboratories that failed to detect her. (For her work with such a robot at M.I.T., she wore a white mask in order to be seen.) The long-term solution for such lapses is “having more folks that look like the United States population at the table when technology is designed,” said Chris S. Crawford, a professor at the University of Alabama who works on direct brain-to-robot controls. Algorithms trained mostly on white male faces (by mostly white male developers who don’t notice the absence of other kinds of people in the process) are better at recognizing white males than other people. “I personally was in Silicon Valley when some of these technologies were being developed,” he said. More than once, he added, “I would sit down and they would test it on me, and it wouldn’t work. And I was like, You know why it’s not working, right?” Robot researchers are typically educated to solve difficult technical problems, not to consider societal questions about who gets to make robots or how the machines affect society. So it was striking that many roboticists signed statements declaring themselves responsible for addressing injustices in the lab and outside it. They committed themselves to actions aimed at making the creation and usage of robots less unjust.

## Perm

### Do both

#### Perm do both – the pitfalls of AI are best solved by political solutions by nation states that utilize existing technologies to create legally binding regulatory policies to combat overambitions

**Sharkey 18** (Noel, Emeritus Professor at the University of Sheffield, 8-28-2018, accessed on 6-25-2022, Humanitarian Law & Policy Blog, "The impact of gender and race bias in AI - Humanitarian Law & Policy Blog", https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/)-qcl

It should be clear from evidence presented above that both AI decision algorithms and face recognition algorithms can be alarmingly biased or inaccurate with darker shades of skin and with women. These may well improve over time but there have been no magic bullet solutions despite massive efforts and several announcements. Many of the companies developing software, particularly for policing, insist that they did well on their inhouse testing. It has remained for other organisations, such as NGOs, to collect the data and demonstrate the biases, yet the systems keep on getting rolled out. It is the familiar old story that once there has been huge investment in a technology it continues to be used despite its failings. Let us not make the same mistake with targeting technology. Discriminatory systems are bad enough in the civilian world where new cases of injustice to women and people with darker shades of skin are turning up almost weekly. But while it can be difficult for those who suspect discrimination to take legal action, there is at least the potential to reverse such unjust decisions. It is a different story when dealing with the technologies of violence. Once someone has been misclassified and targeted with lethal force by an unfairly biased decision process, there is no overturning the decision. Technology, and particularly AI, has always gotten ahead of itself with ambition outstripping achievement. In my long experience working on the subject and reviewing many research proposals, ambition often wins the day. Indeed, ambition is often a positive step towards achievement. In many cases it can still be worthwhile even if the achievement falls well short of the ambition. However, when it comes to technologies of violence, we need to be considerably more cautious of ambitious claims about speculative technology that can lead us down the wrong path. Like a retired police horse, it is time to take off the blinkers and look at the current state of technology and its problematic relationship to the technologies of violence. We cannot simply ignore the types of discriminatory algorithmic biases appearing in the civilian world and pretend that we can just make them go away when it comes weapons development and use. These are just some of the problems that have come to light, since the increased use of AI in society. We don’t know what further problems are around the corner or what further biases are likely to occur in targeting technologies. The moral of this tale is simple. We must take a precautionary approach to the use of AI in weapons technology and AWS in particular. We must not rely on the possibility of future fixes but instead make decisions based on what the technology is capable of today. It is time now for nation States to step up to the mark and begin negotiations for a new international legally binding instrument to ensure the meaningful human control of weapons systems is preserved.

### Bias

#### Perm We will never get perfect bias mitigation – but most effective approach is all-encompassing

UNIDIR (United Nations Institute for Disarmament Research), 2018, “Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies – A Primer”, The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to a variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and Governments. UNIDIR activities are funded by contributions from Governments and donor foundations, <https://www.unidir.org/sites/default/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf>, - Maren Lien

As algorithms approach ubiquity, there is growing understanding that they are not objective and infallible. Algorithms in **all** domains, including military applications, can exhibit multiple types of **biases** that arise from **different sources**, such as unrepresentative training data or inappropriate transfer of the algorithm to a novel context. Some degree of **algorithmic bias** may be **inevitable**, as it might not be possible to satisfy all relevant norms with a single process, decision, or algorithm. At the same time, algorithmic biases are not mutually exclusive, as some biases feed into one another. Moreover, not all biases are bad, as some biases can be beneficial to achieving the user’s end goals. Most pointedly, **algorithmic** **bias** can arise at every stage of **development** and **deployment**, with each stage bringing its own set of considerations and possibilities for the outcome of bias. In many cases, **mitigation** strategies are **available**, but they require **careful** engagement with the details of the **situation**, as one might not want to mitigate; or might be able to mitigate only some biases; or might address problems by changing the users or broader system; and so forth. Various institutions and organizations are beginning to address these challenges, though policy and technical responses are still in their infancy. As a contribution to the policy response, those participating in the discussion on LAWS within the CCW framework may wish to consider the following questions about algorithmic biases in future systems: • If governments decide to regulate increasingly autonomous weapon systems, rather than adopt an outright ban, which national or international organizations or instruments would be best placed to offer guidance or assistance to address potential algorithmic biases in AWS, including identifying possible mitigation steps? • Given the secretive or non-transparent nature of weapon development and weapon review processes, what sorts of “best practices” can provide confidence that key algorithmic biases have been appropriately identified and mitigated? • Are mitigation steps for algorithmic biases in particular AWS robust against possible loss of communication, interoperability challenges, or reduced human oversight? • How would **training** of **operators** and **commanders** need to be **adapted** to ensure that they appropriately understand the **algorithmic biases** in an **AWS**, in order to maintain **trust** in the **system** and ensure its **lawful** use?

### US leadership

#### Perm - US is most favorable to ethical principles for AI weapons

David H. Freedman, 9/15/21, “US Is Only Nation with Ethical Standards for AI Weapons. Should We Be Afraid?”, Newsweek, David H. Freedman is a scientific journalist, author, and is a contributing writer at The Atlantic and Newsweek, <https://www.newsweek.com/2021/09/24/us-only-nation-ethical-standards-ai-weapons-should-we-afraid-1628986.html> - Maren Lien

Even if military AI systems work exactly as intended, is it ethical to give machines the authority to destroy and kill? Work, the former defense deputy secretary, insists the **U.S.** **military** is strictly committed to keeping a **human** **decision-maker** in the "**kill chain**" so that no **weapon** will pick a target and **fire on its own** without an OK. But other nations may not be as careful, he says. "As far as we know, the **U.S. military** is the **only one** that has established **ethical** principles for **AI." Twenty-two nations** have asked the **U**nited **N**ations to **ban** automated weapons capable of operating **outside human oversight**, but so far **no agreements** have been signed. Human Rights Watch and other advocacy groups have called for similar bans to no avail.

#### Perm True mitigation requires multiple sets of actors – including international government regulators

UNIDIR (United Nations Institute for Disarmament Research), 2018, “Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies – A Primer”, The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to a variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and Governments. UNIDIR activities are funded by contributions from Governments and donor foundations, <https://www.unidir.org/sites/default/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf>, - Maren Lien

Responsibility for **mitigating** unwanted **algorithmic biases** does **not** rest with a **single actor**. A first set of **actors** are the **program developers** designing and creating the system. The developer is intimately familiar with each of the algorithms running in the system. To the extent that an undesirable bias can be mitigated through changes in the underlying algorithms or development process, then developers present a natural locus of intervention. In this way, some potential problems can be avoided before the system is fully built. At the same time, not all algorithmic biases can be addressed purely in the development stage. For example, appropriate training data might not be available, and the developers might have insufficient knowledge of deployment contexts to appropriately adjust their algorithms. The second set of key actors in potential **mitigation** of AWS algorithm biases are the **acquirers** of the technology. The agency or organization responsible for the purchase of the technology can require that the system have certain features, or meet specific, pre-defined standards. Alternately, the acquirer can require that the developers provide them with precise, detailed information about the training data, intended use contexts, and so forth. In the former case, the acquirer indicates which algorithmic biases are unacceptable, and the developer must find some way of producing such a system. In the latter case, the acquirer gains the knowledge needed to adapt practices (such as rules of engagement) to minimize the harms from the algorithmic biases that remain. In either case, acquisition and procurement teams can minimize the likelihood of algorithmic “failures” or negative biases. The third set of potential actors in **mitigation** efforts are **regulators** (including **international policymakers**) and testers. Regulators could decide to **completely ban** the **development** or **use** of **AWS**. Alternatively, they may decide to restrict or regulate some facet of development or use. In this case, they may determine which algorithmic **biases** are **unacceptable**, and not allow **deployment** of systems that **exhibit** those **biases**. They could prioritize various conditions, properties, and behaviours of a weapon system, and thereby impose particular **ethical**, **legal**, or **social norms** that the **system** must **follow**, though the developers are left with the task of determining how to satisfy those constraints. **National** or **international regulators** also have the ability to **dictate regulatory constraints and processes** that can help guide developers and future testers in their search for these or similar-acting system biases. Lastly, through testing, some algorithmic biases may be identified prior to approval and deployment, allowing for system revisions prior to the negative, real-world or real-life impacts that would impair efficacy or trust in future AWS deployment. The fourth set of potential actors would be the deployers or operators of the system. These actors, whether at the strategic or tactical level, would make the final decisions about whether, when and where to use the weapon system, and so have the ability to mitigate algorithmic biases simply by not using the system. Alternately, if a system is used only in settings for which it was designed with appropriate training data (and all of the other conditions), then the system’s potentially harmful impacts will be mitigated—though not necessarily completely eliminated.

### Governance

#### AI can be made ethical through policy and collaboration

Blackman 20 Harvard Business Review, 10-15-2020, "A Practical Guide to Building Ethical AI," <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai> [AJL]

How to Operationalize Data and AI Ethics AI ethics does not come in a box. Given the varying values of companies across dozens of industries, a data and AI ethics program must be tailored to the specific business and regulatory needs that are relevant to the company. However, here are seven steps towards building a customized, operationalized, scalable, and sustainable data and AI ethics program. 1. Identify existing infrastructure that a data and AI ethics program can leverage. The key to a successful creation of a data and AI ethics program is using the power and authority of existing infrastructure, such as a data governance board that convenes to discuss privacy, cyber, compliance, and other data-related risks. This allows concerns from those “on the ground” (e.g., product owners and managers) to bubble up and, when necessary, they can in turn elevate key concerns to relevant executives. Governance board buy in works for a few reasons: 1) the executive level sets the tone for how seriously employees will take these issues, 2) a data and AI ethics strategy needs to dovetail with the general data and AI strategy, which is devised at the executive level, and 3) protecting the brand from reputational, regulatory, and legal risk is ultimately a C-suite responsibility, and they need to be alerted when high stakes issues arise. If such a body does not exist then companies can create one — an ethics council or committee, for example — with ethics-adjacent personnel, such as those in cyber, risk and compliance, privacy, and analytics. It may also be advisable to include external subject matter experts, including ethicists. 2. Create a data and AI ethical risk framework that is tailored to your industry. A good framework comprises, at a minimum, an articulation of the ethical standards — including the ethical nightmares — of the company, an identification of the relevant external and internal stakeholders, a recommended governance structure, and an articulation of how that structure will be maintained in the face of changing personnel and circumstances. It is important to establish KPIs and a quality assurance program to measure the continued effectiveness of the tactics carrying out your strategy. A robust framework also makes clear how ethical risk mitigation is built into operations. For instance, it should identify the ethical standards data collectors, product developers, and product managers and owners must adhere to. It should also articulate a clear process by which ethical concerns are elevated to more senior leadership or to an ethics committee. All companies should ask whether there are processes in place that vet for biased algorithms, privacy violations, and unexplainable outputs. Still, frameworks also need to be tailored to a company’s industry. In finance, it is important to think about how digital identities are determined and how international transactions can be ethically safe. In health care there will need to be extra protections built around privacy, particularly as AI enables the development of precision medicine. In the retail space, where recommendation engines loom large, it is important to develop methods to detect and mitigate associative bias, where recommendations flow from stereotypical and sometimes offensive associations with various populations. 3. Change how you think about ethics by taking cues from the successes in health care. Many senior leaders describe ethics in general — and data and AI ethics in particular — as “squishy” or “fuzzy,” and argue it is not sufficiently “concrete” to be actionable. Leaders should take inspiration from health care, an industry that has been systematically focused on ethical risk mitigation since at least the 1970s. Key concerns about what constitutes privacy, self-determination, and informed consent, for example, have been explored deeply by medical ethicists, health care practitioners, regulators, and lawyers. Those insights can be transferred to many ethical dilemmas around consumer data privacy and control. For instance, companies attest to respect the users of their products, but what does that mean in practice? In health care, an essential requirement of demonstrating respect for patients is that they are treated only after granting their informed consent — understood to include consent that, at a minimum, does not result from lies, manipulation, or communications in words the patient cannot understand, such as impenetrable legalese or Latin medical terms. These same kinds of requirements can be brought to bear on how people’s data is collected, used, and shared. Ensuring that users are not only informed of how their data is being used, but also that they are informed early on and in a way that makes comprehension likely (for instance, by not burying the information in a long legal document), is one easy lesson to take from health care. The more general lesson is to break down big ethical concepts like privacy, bias, and explainability into infrastructure, process, and practice that realize those values. 4. Optimize guidance and tools for product managers. While your framework provides high-level guidance, it’s essential that guidance at the product level is granular. Take, for instance, the oft-lauded value of explainability in AI, a highly valued feature of ML models that will likely be part of your framework. Standard machine-learning algorithms engage in pattern recognition too unwieldy for humans to grasp. But it is common — particularly when the outputs of the AI are potentially life-altering — to want or demand explanations for AI outputs. The problem is that there is often a tension between making outputs explainable, on the one hand, and making the outputs (e.g. predictions) accurate, on the other. Product managers need to know how to make that tradeoff, and customized tools should be developed to help product managers make those decisions. For example, companies can create a tool by which project managers can evaluate the importance of explainability for a given product. If explainability is desirable because it helps to ferret out bias in an algorithm, but biased outputs are not a concern for this particular ML application, then that downgrades the importance of explainability relative to accuracy. If the outputs fall under regulations that require explanations — for instance, regulations in the banking industry that require banks to explain why someone has been turned down for a loan — then explainability will be imperative. The same goes for other relevant values, e.g. which, if any, of the dozens of metrics to use when determining whether a product delivers fair or equitable outputs. 5. Build organizational awareness. Ten years ago, corporations scarcely paid attention to cyber risks, but they certainly do now, and employees are expected to have a grasp of some of those risks. Anyone who touches data or AI products — be they in HR, marketing, or operations — should understand the company’s data and AI ethics framework. Creating a culture in which a data and AI ethics strategy can be successfully deployed and maintained requires educating and upskilling employees, and empowering them to raise important questions at crucial junctures and raise key concerns to the appropriate deliberative body. Throughout this process, it’s important to clearly articulate why data and AI ethics matters to the organization in a way that demonstrates the commitment is not merely part of a public relations campaign. 6. Formally and informally incentivize employees to play a role in identifying AI ethical risks. As we’ve learned from numerous infamous examples, ethical standards are compromised when people are financially incentivized to act unethically. Similarly, failing to financially incentivize ethical actions can lead to them being deprioritized. A company’s values are partly determined by how it directs financial resources. When employees don’t see a budget behind scaling and maintaining a strong data and AI ethics program, they will turn their attention to what moves them forward in their career. Rewarding people for their efforts in promoting a data ethics program is essential. 7. Monitor impacts and engage stakeholders. Creating organizational awareness, ethics committees, informed product managers owners, engineers, and data collectors is all part of the development and, ideally, procurement process. But due to limited resources, time, and a general failure to imagine all the ways things can go wrong, it is important to monitor the impacts of the data and AI products that are on the market. A car can be built with air bags and crumple zones, but that doesn’t mean it’s safe to drive it at 100 mph down a side street. Similarly, AI products can be ethically developed but unethically deployed. There is both qualitative and quantitative research to be done here, including especially engaging stakeholders to determine how the product has affected them. Indeed, in the ideal scenario, relevant stakeholders are identified early in the development process and incorporated into an articulation of what the product does and does not do. Operationalizing data and AI ethics is not an easy task. It requires buy-in from senior leadership and cross-functional collaboration. Companies that make the investment, however, will not only see mitigated risk but also more efficient adoption of the technologies they need to forge ahead. And finally, they’ll be exactly what their clients, consumers, and employees are looking for: trustworthy.

### Distributed cognition

#### **Distributed cognitions models enhance human life – post humanist ideologies help us fashion new modes of existence and explore the potential of virtual technologies**

Hayles ’99 (N. Katherine Hayles is a literary critic and theorist. She is the author of *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics* which won the Rene Wellek Prize for the best book in literary theory for 1998–1999.--“How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics”—Feb 15, 1999—book--<https://monoskop.org/images/5/50/Hayles_N_Katherine_How_We_Became_Posthuman_Virtual_Bodies_in_Cybernetics_Literature_and_Informatics.pdf)//Marzz>

Hutchins would no doubt disagree with Weizenbaum's view that judgment should be reserved for humans alone. Like cognition, decision making is distributed between human and nonhuman agents, from the steam-powered steering system that suddenly failed on a navy vessel Hutchins was studying to the charts and pocket calculators that the navigators were then forced to use to calculate their position. He convincingly shows that these adaptations to changed circumstances were evolutionary and embodied rather than abstract and consciously designed (pp. 347-51). The solution to the problem caused by this sudden failure of the steering mechanism was" clearly discovered by the organization [of the system as a whole] before it was discovered by any of the participants" (p. 361). Seen in this perspective, the prospect of humans working in partnership with intelligent machines is not so much a usurpation of human right and responsibility as it is a further development in the construction of distributed cognition environments, a construction that has been ongoing for thousands of years. Also changed in this perspective is the relation of human subjectivity to its environment. No longer is human will seen as the source from which emanates the mastery necessary to dominate and control the environment. Rather, the distributed cognition of the emergent human subject correlates with-in Bateson's phrase, becomes a metaphor for-the distributed cognitive system as a whole, in which "thinking" is done by both human and nonhuman actors. "Thinking consists of bringing these structures into coordination so they can shape and be shaped by one another," Hutchins wrote (p. 316). To conceptualize the human in these terms is not to imperil human survival but is precisely to enhance it, for the more we understand the flexible, adaptive structures that coordinate our environments and the metaphors that we ourselves are, the better we can fashion images of ourselves that accurately reflect the **complex interplays** that ultimately make the entire world one system. This view of the posthuman also offers resources for thinking in more sophisticated ways about virtual technologies. As long as the human subject is envisioned as an autonomous self with unambiguous boundaries, the human-computer interface can only be parsed as a division between the solidity of real life on one side and the illusion of virtual reality on the other, thus obscuring the far-reaching changes initiated by the development of virtual technologies. Only if one thinks of the subject as an autonomous self independent of the environment is one likely to experience the panic performed by Norbert Wiener's Cybernetics and Bernard Wolfe's Limbo. This view of the self authorizes the fear that if the boundaries are breached at all, there will be nothing to stop the self's complete dissolution. By contrast, when the human is seen as part of a distributed system, the full expression of human capability can be seen precisely to depend on the splice rather than being imperiled by it. Writing in another context, Hutchins arrives at an in Sight profoundly applicable to virtual technologies: "What used to look like internalization [of thought and subjectivity] now appears as a gradual propagation of organized functional properties across a set of malleable media" (p. 312). This vision is a potent antidote to the view that parses virtuality as a division between an inert body that is left behind and a disembodied subjectivity that inhabits a virtual realm, the construction of virtuality performed by Case in William Gibson's Neuromancer when he delights in the "bodiless exultation of cyberspace" and fears, above all, dropping back into the "meat" of the body.22 By contrast, in the model that Hutchins presents and that the posthuman helps to authorize, human functionality expands because the parameters of the cognitive system it inhabits expand. In this model, it is not a question of leaving the body behind but rather of extending embodied awareness in highly specific, local, and material ways that would be impossible without electronic prosthesis.

### LAWS

#### LAWS produce ethical and logistic quandaries – weaponizing AI accelerates causalities and human rights abuses

Amnesty International 15 (Amnesty International is a global movement of more than 3 million supporters, members and activists in more than 150 countries and territories who campaign to end grave abuses of human rights. Our vision is for every person to enjoy all the rights enshrined in the Universal Declaration of Human Rights and other international human rights standards. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and public donations.) “AUTONOMOUS WEAPONS SYSTEMS: FIVE KEY HUMAN RIGHTS ISSUES FOR CONSIDERATION” April 10 2015 <https://www.amnesty.org/en/documents/act30/1401/2015/en/> // ZX

Over the past decade, there have been extensive advances in artificial intelligence and other technologies. These will make possible the development and deployment of fully autonomous weapons systems which, once activated, can select, attack, kill and wound human targets, and will be able to operate without effective human control. These weapons systems are often referred to as Lethal Autonomous Robotics (LARs), Lethal Autonomous Weapons Systems (LAWS) and, more comprehensively, Autonomous Weapons Systems (AWS). The rapid development of these weapons systems could not only change the entire nature of warfare, it could also dramatically alter the conduct of law enforcement operations and raises extremely serious human rights concerns, undermining the right to life, the prohibition of torture and other ill-treatment, and the right to security of person, and other human rights. Amnesty International has taken the view that AWS is a useful term for these weapons systems, since these systems can (i) be designed to have lethal or less lethal effects and (ii) be used in armed conflict and/or law enforcement situations. With proliferation they are likely to come to be used by non-state armed groups, criminal gangs and private companies and individuals. Amnesty International takes the term ‘autonomous’ to mean weapons capable of selecting targets and triggering an attack without effective or meaningful human control1 that can ensure the lawful use of force. Such systems would use violence (including less-lethal force) against individuals, and could have adverse consequences for a person’s human rights. While the development of AWS clearly raises serious and legitimate ethical and societal concerns, this briefing paper will examine the implications of AWS in the context of international law, particularly international human rights law and standards. The important concerns around their use in situations of armed conflict, and thus their ability to comply fully with international humanitarian law (IHL), has been the focus of previous work on AWS, including by Human Rights Watch, other members of the Campaign to Stop Killer Robots and the International Committee of the Red Cross (ICRC). This briefing paper, however, will address some of the implications for human rights related to AWS, particularly those rights and standards that govern the conduct of law enforcement operations. Amnesty International believes that the questions surrounding the development and potential use of AWS outside armed conflict (and the ability of such systems to comply with human rights law) are at least as daunting as those related to their use on the battlefield and urgently require attention and consideration2 , ultimately leading to concrete steps that will address this important area of international law. Amnesty International has identified five key human rights issues for consideration in the current debate on AWS: 1) The scope of the Convention on Certain Conventional Weapons (CCW) does not cover non-conflict situations; 2) AWS will not be able to comply with relevant international human rights law (IHRL) and policing standards; 3) Developments in existing semi-autonomous weapons technology pose fundamental challenges for the IHRL framework; 4) In the absence of a prohibition, AWS must be subject to independent weapons reviews; and 5) AWS will erode accountability mechanisms. The issues identified are by no means exhaustive, but rather seek to elucidate the principal concerns around the potential use of AWS in law enforcement operations. This briefing argues that the use of AWS, including less-lethal robotic weapons, in law enforcement operations would be fundamentally incompatible with international human rights law, and would lead to unlawful killings, injuries and other violations of human rights. Furthermore, the use of AWS would pose serious challenges in holding accountable those responsible for serious violations and could entrench impunity for crimes under international law. Consequently, Amnesty International supports the call for a pre-emptive ban on the development, transfer, deployment and use of AWS, including fully autonomous systems that deploy less-lethal weapons and can result in death or serious injury. In the absence of a prohibition, Amnesty International supports the call of UN Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, to impose a moratorium on the development, transfer, deployment and use of AWS and ensure that moratorium covers both lethal and less-lethal weapons. This principle deals with two different thresholds: a) when it is appropriate to use firearms (potentially lethal force) and b) the even higher threshold of when the intentional lethal use of firearms is permissible. Each of these situations involves a complex assessment of potential or imminent threats to life or serious injury and how to respond to them appropriately, and it involves deciding how best to protect the right to life, which is an absolutely fundamental duty of the state under human rights law. Such life and death decisions must never be delegated to AWS. In order to be able to carry out policing and law enforcement operations in a lawful manner, AWS would need to be able to effectively assess the degree to which there was an imminent threat of death or serious injury, identify correctly who is posing the threat, consider whether force is necessary to neutralize the threat, be able to identify and use means other than force, have the capacity to deploy different modes of communication and policing weapons and equipment to allow for a graduated response, and have available back up means and resources. To add to this complexity, each situation would require a different and unique response, which would be extremely challenging to reduce to a series of complex algorithms. It is not possible that AWS, without meaningful and effective human control and judgement, would be able to comply with these provisions, especially in unpredictable and ever-evolving environments. In an open letter in October 2013, computer scientists, engineers, artificial intelligence experts, roboticists and professionals from related disciplines from 37 countries asserted that “in the absence of clear scientific evidence that robot weapons have, or are likely to have in the foreseeable future, the functionality required for accurate target identification, situational awareness or decisions regarding the proportional use of force, we question whether they could meet the strict legal requirements for the use of force” and that “[G]iven the limitations and unknown future risks of autonomous robot weapons technology…,[D]ecisions about the application of violent force must not be delegated to machines.”15 The UNBPUFF places a due diligence requirement upon states to review weapons used in law enforcement. As Principle 3 of the UNBPUFF states, “the development and deployment of non-lethal incapacitating weapons should be carefully evaluated in order to minimize the risk of endangering uninvolved persons”. This review is limited to less-lethal weapons but is still important to ensure that those weapons will comply with relevant international standards and national laws and, moreover, given that evidence shows that “non-lethal” weapons can often have lethal effects which is why the term “less-lethal” is more appropriate. The requirement of a review of weapons used for law enforcement is even more important given the increasing ‘militarization’ of law enforcement operations, whereby military personnel assume roles often held by law enforcement agencies, such as policing of public assemblies. In the absence of a prohibition on AWS, states intending to develop, acquire, or use AWS must therefore be required to thoroughly review whether they can be used in a manner that fully respects relevant law and standards be it for law enforcement or military operations. This testing should be carried out by an independent body. The rapid technological advances that are moving towards full autonomy in weapons systems present serious concerns. The technology to allow fully autonomous operations may be reached soon; but it is extremely unlikely that programming that could ensure AWS perform law enforcement functions lawfully would be developed in the foreseeable future. Any new law enforcement equipment should be introduced based on clearly defined operational needs and technical requirements with a view to reduce the amount of force used and the risk and level of harm and injury caused. They must be subject to rigorous testing, by an independent expert body, and the testing, review and selection process should be legally constituted. In addition to assessing compliance with the UNBPUFF themselves, the process must test AWS compatibility with other key human rights treaties and standards, including ICCPR, International Covenant on Economic, Social and Cultural Rights (CESCR), the Convention Against Torture, the SMRTP and the UNCCLEO.

## AI Turns

#### Companies use AI in a way that has a significant impact on people’s lives

Kearns and Roth 19 Kearns, Michael, and Aaron Roth. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Illustrated, Oxford University Press, 2019. [AJL]

Which all brings us to a conundrum. The insights we can get from this unprecedented access to data can be a great thing: we can get new understanding about how our society works, and improve public health, municipal services, and consumer products. But as individuals, we aren’t just the recipients of the fruits of this data analysis: we are the data, and it is being used to make decisions about us—sometimes very consequential decisions. In December 2018, the New York Times obtained a commercial dataset containing location information collected from phone apps whose nominal purpose is to provide mundane things like weather reports and restaurant recommendations. Such datasets contain precise locations for hundreds of millions of individuals, each updated hundreds (or even thousands) of times a day. Commercial buyers of such data will generally be interested in aggregate information—for example, a hedge fund might be interested in tracking the number of people who shop at a particular chain of retail outlets in order to predict quarterly revenues. But the data is recorded by individual phones. It is superficially anonymous, without names attached—but there is only so much anonymity you can promise when recording a person’s every move. For example, from this data the New York Times was able to identify a forty-six-year-old math teacher named Lisa Magrin. She was the only person who made the daily commute from her home in upstate New York to the middle school where she works, fourteen miles away. And once someone’s identity is uncovered in this way, it’s possible to learn a lot more about them. The Times followed Lisa’s data trail to Weight Watchers, to a dermatologist’s office, and to her ex-boyfriend’s home. She found this disturbing and told the Times why: “It’s the thought of people finding out those intimate details that you don’t want people to know.” Just a couple of decades ago, this level of intrusive surveillance would have required a private investigator or a Introduction ■ 3 government agency; now it is simply the by-product of widely available commercial datasets. Clearly, we have entered a brave new world. And it’s not only privacy that has become a concern as data gathering and analysis proliferate. Because algorithms—those little bits of machine code that increasingly mediate our behavior via our phones and the Internet—aren’t simply analyzing the data that we generate with our every move. They are also being used to actively make decisions that affect our lives. When you apply for a credit card, your application may never be examined by a human being. Instead, an algorithm pulling in data about you (and perhaps also about people “like you”) from many different sources might automatically approve or deny your request. Though there are benefits to knowing instantaneously whether your request is approved, rather than waiting five to ten business days, this should give us a moment of pause. In many states, algorithms based on what is called machine learning are also used to inform bail, parole, and criminal sentencing decisions. Algorithms are used to deploy police officers across cities. They are being used to make decisions in all sorts of domains that have direct and real impact on people’s lives. All this raises questions not only of privacy but also of fairness, as well as a variety of other basic social values including safety, transparency, accountability, and even morality. So if we are going to continue to generate and use huge datasets to automate important decisions (a trend whose reversal seems about as plausible as our returning to an agrarian society), we have to think seriously about some weighty topics. These include limits on the use of data and algorithms, and the corresponding laws, regulations, and organizations that would determine and enforce those limits. But we must also think seriously about addressing the concerns scientifically—about what it might mean to encode ethical principles directly into the design of the algorithms that are increasingly 4 ■ THE ETHICAL ALGORITHM woven into our daily lives. This book is about the emerging science of ethical algorithm design, which tries to do exactly that.

#### AI can help in international peace efforts – no existential threat from AI

**Daanish** Masood and **Martin** Waehlisch, 2019-04-23, “AI & Global Governance: Robots Will Not Only Wage Future Wars also Future Peace, United Nations University Centre for Policy Research**, Daanish Masood and Martin Waehlisch are Political Affairs Officers at the UN’s Department of Political and Peacebuilding Affairs.** <https://cpr.unu.edu/publications/articles/robots-will-not-only-wage-future-wars-but-also-future-peace.html> **- Maren Lien**

Though touted as a real possibility by the likes of Elon Musk, that particular idea has been dismissed in the field as far-fetched. In his 2018 book, Ten Arguments for Deleting Your Social Media Accounts Now, polymathic computer scientist and ‘founding father’ of virtual reality Jaron Lanier described AI as a decades-old lie that he and others in Silicon Valley invented just to get money from DARPA, the US Pentagon agency responsible for researching technological breakthroughs. Lanier was being tongue-in-cheek. His point was that despite our dystopian fears, **AI** is still far too **rudimentary** to pose an **existential threat** to the **human species**. At the United Nations, we have been exploring completely different scenarios for AI: its **potentia**l to be used for the noble purposes of **peace and security**. This could **revolutionize** the way of how we **prevent** and **solve conflicts globally**. Two of the most promising areas are Machine Learning and Natural Language Processing. Machine Learning involves computer algorithms detecting patterns from data to learn how to make predictions and recommendations. Natural Language Processing involves computers learning to understand human languages. At the UN Secretariat, our chief concern is with how these emerging technologies can be deployed for the good of humanity to de-escalate violence and increase international stability. This endeavor has admirable precedent. During the **Cold War**, computer scientists used **multilayered** simulations to predict the **scale** and **potential** outcome of the arms race between the East and the West. Since then, governments and international agencies have **increasingly** used **computational** **models** and advanced Machine Learning to try to understand recurrent conflict patterns and forecast moments of state fragility. But two things have transformed the scope for progress in this field. The first is the sheer volume of data now available from what people say and do online. The second is the game-changing growth in computational capacity that allows us to crunch unprecedented, inconceivable quantities data with relative speed and ease. So how can this help the United Nations build peace? Three ways come to mind. Firstly, overcoming cultural and language barriers. By teaching **computers** to understand human language and the nuances of dialects, not only can we better link up what people write on social media to local contexts of conflict, we can also more methodically follow what people **say** on radio and TV. As part of the UN’s early warning efforts, this can help us **detect hate speech** in a place where the potential for **conflict** is high. This is crucial because the UN often works in countries where internet coverage is low, and where the spoken languages may not be well understood by many of its international staff. Natural Language Processing algorithms can help to **track** and improve understanding of **local debates**, which might well be **blind spots** for the international community. If we combine such methods with Machine Learning chatbots, the UN could conduct large-scale digital focus groups with thousands in real-time, enabling different demographic segments in a country to voice their views on, say, a proposed peace deal – instantly testing public support, and indicating the chances of sustainability. Secondly, anticipating the deeper drivers of conflict. We could combine new imaging techniques – whether satellites or drones – with automation. For instance, many parts of the world are experiencing severe groundwater withdrawal and water aquifer depletion. Water scarcity, in turn, drives conflicts and undermines stability in post-conflict environments, where violence around water access becomes more likely, along with large movements of people leaving newly arid areas. One of the best predictors of water depletion is land subsidence or sinking, which can be measured by satellite and drone imagery. By combining these **imaging techniques** with **Machine Learning**, the UN can work in **partnership** with **governments and local communities** to **anticipate** future **water conflicts** and begin working **proactively** to **reduce** their likelihood. Thirdly, advancing decision making. In the work of peace and security, it is surprising how many consequential decisions are still made solely on the basis of intuition. Yet **complex** decisions often need to navigate conflicting goals and undiscovered options, against a landscape of **limited information** and political preference. This is where we can use Deep Learning – where a **network** can absorb **huge** amounts of public **data** and **test** it against **real-world** examples on which it is trained while applying with probabilistic modeling. This mathematical approach can help us to generate models of our uncertain, dynamic world with limited data. With better data, we can eventually make better predictions to guide **complex** decisions. Future senior peace envoys charged with mediating a conflict would benefit from such advances to stress test elements of a peace agreement. Of course, **human decision**-making will remain crucial, but would be **informed** by more **evidence-driven** robust **analytical** tools. Doing the above inside the UN, will require training staff and senior leaders in new approaches and trusting in their competence. And it will also require collaborating with university researchers, and forging close partnerships with leading private AI and technology firms. The good news is that the work has already started. But we are still at baby-steps. With the Secretary-General’s support, including through his landmark Strategy on New Technologies, the time to scale this activity has come. We can leave no stone unturned and no tool ignored to reduce violence and promote peace – that, after all, is the moral obligation at the very core of the UN Charter.

## AT Alt

### Ivory tower DA

#### Academic alts fail – they don’t make useable alts

Reid Blackman, October 15 2020, “A Practical Guide to Building Ethical AI”, Harvard Business Review, Reid Blackman, Ph.D., is the author of the book Ethical Machines (Harvard Business Review Press, July 2022) and Founder and CEO of Virtue, an AI ethical risk consultancy. He has also been a Senior Advisor to the Deloitte AI Institute, a Founding Member of Ernst & Young’s AI Advisory Board, and volunteers as the Chief Ethics Officer to the non-profit Government Blockchain Association. Reid’s expertise is relied upon by Fortune 500 and Global 1000 companies to speak to and educate their people and to guide them as they create and scale AI ethical risk programs., <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai>

First, there is the **academic approach**. **Academics** — and I speak from 15 years of experience as a former professor of philosophy — are fantastic at **rigorous** and **systematic** inquiry. Those academics who are ethicists (typically found in philosophy departments) are adept at spotting ethical problems, their sources, and how to think through them. But while academic ethicists might seem like a perfect match, given the need for systematic identification and mitigation of ethical risks, they unfortunately tend to ask **different** questions than businesses. For the most part, academics ask, “Should we do this? Would it be good for society overall? Does it conduce to human flourishing?” Businesses, on the other hand, tend to ask, “Given that we are going to do this, how can we do it without making ourselves vulnerable to ethical risks?” The result is academic treatments that **do not** speak to the **highly particular**, **concrete** uses of **data and AI**. This translates to the **absence** of **clear directives** to the developers on the ground and the senior leaders who need to identify and choose among a set of **risk mitigation** strategies.