The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines

Ash Carter

This essay takes an engineering approach to ensuring that the deployment of artificial intelligence does not confound ethical principles, even in sensitive applications like national security. There are design techniques in all three parts of the AI architecture – algorithms, data sets, and applications – that can be used to incorporate important moral considerations. The newness and complexity of AI cannot therefore serve as an excuse for immoral outcomes of deployment by companies or governments.

One of the most frequent questions I was asked as U.S. Secretary of Defense (2015 – 2017) was whether there will be autonomous lethal weapons. My answer was no, the U.S. Department of Defense (DOD) would not deploy or use truly autonomous systems in the application of lethal force. Being technologically inclined, I established the Pentagon's official policy in a memorandum back in 2012 when I was Deputy Secretary. When conceiving of this directive, I had imagined myself standing in front of the news cameras the morning after innocent bystanders had been killed in an airstrike aimed at terrorists or opposing combatants. And suppose I answered in response to obvious and justified interrogation over responsibility: "It's tragic, but it's not our fault: the machine did it." This reply would be rightly regarded as unacceptable and immoral.

What, then, can ethically "justify" the risk of a terrible error made in the application of artificial intelligence?[1] In one sense, nothing, of course. Yet as a practical matter, AI is going to be used, and in an ever-widening set of applications. So what can bound moral error? Algorithm design? Data set selection and editing? Restricting or even banning use in sensitive applications? Diligent, genuine, and documented efforts to avoid tragedies? To some extent, all of these.[2] The fact that there are practical technical approaches to responsible use of AI is paramount to national defense. AI is an important ingredient of the necessary transformation of the U.S. military's armamentarium to the greater use of new technologies, almost all of them AI-enabled in some way.

https://doi.org/10.1162/DAED_a_01917

This essay takes a technical rather than a legal approach to AI ethics. It explores some practical methods to minimize and explain ethical errors. It provides some reasons to believe that the good to be obtained by deployment of AI can far outweigh the ethical risks.

The 2012 DOD Directive 3000.09 reads "autonomous and semi-autonomous weapons systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force."[3] This guidance has been reissued and refined several times since.

Note that the directive does *not* use the language "man in the loop." To use such a formulation would be technically ignorant and utterly infeasible. The whole point of the machine is to operate faster, more accurately, and frequently entirely without communication with humans (that is, "autonomously"). Thus, the image of a person inserted in the circuitry like a living chip is ridiculous. In certain ways, the whole idea of autonomy in warfare is not at all new. Take a guided anti-air missile, for example: most of these find their way to their target – during their entire trajectory, or at least in their lethal "endgame" – using inputs from a homing seeker in the nose (a camera, say, or radar) whose output is calculated on board the missile with computers and software designed and tested years previously and updated during flight. In these respects, the question about autonomous weapons or "lethal AI" has thus been around for quite a while. Still, with AI and many of its applications developing at lightning speed, we must give some good answers to its distinctive questions. The language of the directive was crafted to suggest that the DOD would insist upon other more practical forms of "human judgment" built into its AI-enabled weapons systems.

It is not particularly surprising that the tradecraft for ethical use of AI has lagged development of the technology itself. For one thing, after pioneering AI as early as the 1950s, the DOD ended up lagging in its application to military problems (in a manner all too common). It is relatively recent that the Pentagon has begun catching up. And while universities do critical fundamental research, including interdisciplinary work bringing tech together with law, ethics, policy, and other fields of thought, they are remote from direct application at scale. Instead, applied work in AI has been led by the consumer Internet and advertising industries. These industries can afford to be tolerant of Type I errors (pushing content to users unlikely to buy the sponsor's products, for example) or Type II errors (not pushing content to likely customers) in accomplishing their objectives. The analogous kinds of errors would be much graver in applications like national security, health care, self-driving vehicles, or law enforcement. The ethical errors of privacy violations and manipulation that have clearly been made by consumer Internet and advertising companies are not errors of AI, but of a basic lack of moral self-scrutiny. In fact, a significant part of the American creative culture, at least in digital technology, has believed that its dynamism springs from its independence from government and virtually any meaningful regulation (strikingly illustrated by Section 230 of the Communications Decency Act of 1996, which provided broad immunity to online platforms from civil liability for content on their platforms).

But in spite of tech's preference for a freewheeling environment and government's deserved reputation for stodginess, public policy is irreplaceable for technological progress. Government is the artery through which not only vital basic research funding flows, but also the rules, norms, and regulations that fortify acceptance and trust by the population of technological progress as something that is a net positive for humanity. Historically, the "disruptive" technology-enabled industrial revolution that resulted in the gigantic farm-to-factory migration was only successful in America because the government supplied the complementary ingredients to those supplied by innovators and profit-seeking industry. Government-supplied ingredients like universal public education prepared a farmland workforce for industrial jobs. Progressive-era labor reforms offering legal and other safeguards for workers made it possible for most Americans to support a free market system of large corporations and to view technology as a net positive. In other countries, notably Russia, the farm-to-factory revolution failed. During roughly the same period, U.S. government standards for the safety of foods and drugs were promulgated – and accepted – at a nationwide scale. Without such regulation, citizens would not be able to trust the industrialized production of foods, and there might not have been continent-wide markets like those that developed quickly in America.

In a similar way, it seems likely that AI and ethics will mix best when tech and government mix well. The purely technical challenges of ethical AI are hard enough. We do not need a failure of government and industry collaboration to be an obstacle to the ethical fielding of AI. As a technologist as well as a government leader, I believe strongly in both the wonders of AI *and* the importance of morality in engineering. I am also optimistic this can be solved, like every other hard problem, with diligent technology-informed effort. This will be essential for national defense.

While some advancements in AI are breathtakingly new, their novelty should not be exaggerated. Right and wrong are certainly not new. The question has been around a long time: what is a "good reason" for the rest of us to soften the penalty for, or excuse entirely, the people who designed and sold the technology that made a tragic error? Various justifications have long been defined in courts, product warranties, corporate disclosure statements, and press conferences. There is even a rule-of-reason that recognizes that no technology is perfect, so "good" only means "good enough." Not just the morality but also the political practicality of deploying AI hinge on some sort of accountability and responsibility engineered into it that is "good enough" for the purpose. Ethical design principles can be identified in all three components of an AI deployment: algorithms, data sets, and applications.

Many kinds of AI algorithms exist in practice, and even more are being developed or hypothesized. They all make enormous numbers of tiny calculations that combine to make overall inferences that cannot be made quickly by humans, have not been recognized by humans, or even perhaps would never be recognized by humans. These computational methods make literal transparency, normally the starting point for ethical accountability, completely impractical. It is usually impossible to "deconvolve" the series of steps leading to the inferences made by AI.

Moreover, like software of nearly any kind, AI algorithms are the product of many hands and many engineers working in many venues over many years. While blame for an unethical outcome can be attributed to the final vendor or end user, even this is unreasonable unless negligence can be shown, which takes us back to the same fundamental dilemmas.

One approach is to make ethics an internal algorithm design criterion from the start. Doing so successfully may require substantial new conceptual invention in its own right, but this can be as exciting for the coding engineers as maximizing any other design feature, especially if value is attributed to it. The federal government, including the DOD, should fund basic research into ethics-by-algorithms, recognizing that companies will underinvest until some terrible wrong occurs. My experience in technology management suggests that the initial specialist refrain "it can't be done" is usually overcome by making the desired innovation a requirement-to-buy or a weighted factor in competitive source selection.

An additional approach is to focus on the *process* of algorithm design rather than the algorithm itself. The history of processes designed to prevent the misuse of nuclear weapons offers a valuable example. Bombs themselves are outfitted with elaborate coded locks to prevent abuse, which could have the gravest consequences. But any repair, movement, or contact – that is, any process in which bombs are handled, moved, repaired, or altered – requires two people rated in the same specialty (the "two-man rule"). Even I, as Secretary of Defense, was not authorized to be alone with a nuclear weapon. These many simultaneous approaches to security policy, some involving design and some involving process, recognized the ineffable variety of possible failure modes and the absolute necessity to prevent every one of them, all in an essentially unending custodianship of tens of thousands of bombs (the half-life of Plutonium-239 is twenty-two thousand years and Uranium-235 is 703 million years). The complexity of this challenge was deepened with the collapse of the Soviet Union, which fragmented the systems that had served to control one of the world's two biggest collections of such weapons. In the 1990s, I ran the Pentagon program created to assist the post-Soviet militaries to protect and reduce the arsenals they inherited, and I was in awe of the way such a combination of design and process methods safeguarded weapons in a totally unforeseen social disintegration.

Programs of established design principles backed up by dual or multiple checkers with equal training and qualifications and redundant safeguards are widely used in complex systems. Establishing such a design process control can not only reduce the likelihood of errors with advanced AI applications but mitigate, at least partially, the liability assigned to innovators if they do occur. It was precisely such a process protocol that was apparently compromised in the famous case of the Boeing 737-Max. Its in-flight controls were reportedly the cause of two back-to-back airline crashes. The Federal Aviation Administration supposedly failed to provide thorough expert checking of the significant changes to in-flight characteristics that occurred when the older 737 was changed to the Max configuration. Among the fatal mistakes was the sacrifice of an established design criterion in the software itself requiring dual redundant sensor inputs to the fly-by-wire flight controls. Due to the COVID epidemic, most people are by now familiar with the Food and Drug Administration's "safety and efficacy" testing that must precede release of a new vaccine. So the notion of requiring a process of qualified review for sensitive products is hardly new and should be the industry standard for AI.

A dilemma arises from proprietary secrecy. A vendor will not want to disclose the inner workings of its algorithms and data sets; these are sensitive for competitive reasons. Given proprietary concerns, it is advantageous to establish industry-wide standards and a level of government involvement in the certification that these standards are being met. Government routinely handles proprietary secrets of competing companies when it serves as a regulator or customer of advanced technology. Government security classification sometimes can be argued to slow the pace of innovation by preventing the free flow of ideas. But in the case of most AI, the preponderance of innovation is centered in companies, and intercompany secrecy is by far and away the bigger barrier to sharing information, the more so as the research frontier has moved out of universities that publish results openly and into industry.

It is worth noting that AI itself can be a powerful tool in certification testing of AI systems whose workings are impossible for humans to fully grasp. The "checking AI" can perform an exhaustive search for oddities in large numbers of input-output runs and thereby identify design defects without unpacking the full mass of layered calculations. In the same way, AI can conduct cyber defense by probing randomly around the victim's attack surface for unidentified holes, simulating the "rat in a maze" attack (to distinguish it from the attacker who exploits exquisite defects discovered in the victim's defenses – the "jewel thief"). This is just a new case of an old pattern in technology and warfare, in which the same invention that creates new dilemmas can also help protect from those very dilemmas. AI-assisted checking of algorithms can also speed up the process of ethical audit so it does not delay deployment.

The next thing to tackle in ethical AI is the data set the algorithm is trained on (if it is machine learning) or otherwise crunches to make recommendations to the human user. Data sets come from a wide variety of huge caches: enterprise business systems, social media, search engines, public data sets, the entire historical corpus of the written word, and Internet of Things (IoT) and sensor data of all kinds. The trickiest sets are "unstructured data": impressively large jumbles of data collected in an incidental manner.

Generally, it really is true: "garbage in, garbage out." Some open-mindedness is needed, however, in the case of AI. Important hints or suggested solutions might come from running on bad data, but they should not be used for making determinations in sensitive applications.

An "ethical audit" of an AI database begins with its provenance. It seems well established that true anonymity cannot be promised: AI is so thoroughly penetrating that individual identity can almost always be unwound. It turns out that the risk of identification goes up in surprising ways when two databases, assembled "anonymously," are combined. There are technical approaches to enhancing privacy and true anonymity in databases used by AI that seem durable. One example is provided by the various forms of "differential privacy" in which fake data are mixed with true data in a quantified way, preserving some privacy but not entirely spoiling the data's use. The Census Bureau uses differential privacy in its data.

It is also clear that "informed consent" is not a good ethical proxy in data collection and exploitation without expert guardrails. Few of us can really understand on our own the full consequences of our consent. A company selling or deploying AI that abuses personal data should not be able to evade responsibility by citing the supposedly informed consent of the victims.

Some data sets are morally questionable from the start, for example those collected in communist China for purposes of dictatorship and control. It is often said that China will outperform the United States in AI because its population of 1.3 billion, or three times the United States', provides a database size advantage. But I am unaware of any design or implementation of AI that is qualitatively better because of a factor of three in data set size. The real difference is the intrusive methods of Chinese data collection and application. China is indeed likely to excel in the AI of totalitarianism, but this is hardly enviable from an ethical perspective.

Assuming that the data sets used in AI are collected ethically to begin with, three features need to be carefully audited for inaccuracies and biases that could lead to morally fateful events when they are deployed. The audit should encompass the training set (in the case of machine learning), the application set, and potential issues in matching the two. As in the case of algorithms, an ethical case can be built on the characteristics of the data themselves and the process by which they are audited.

To my knowledge, there is no substitute for a qualitative examination with a skeptical eye. Is the entire space of possible data points defined and is there a reasonable presence (or understood absence) of points in some corners (such as an edge subset representing a minority)? Is the set examined against a checklist of possible flaws: biased, outdated, or otherwise unrepresentative? How were the data originally tagged? Way back in the provenance of most data sets is a human tagger who originally assigned a location to each point in a dataspace ("is this a dog or cat?"). And again, as in the case of algorithms, AI itself can help work through a proposed data set against a checklist of possible foibles before deployment. Finally, and again as with algorithms, the process of database audit itself can be given ethical standards: documentation, multiple qualified checkers, simulations, and sampling.

The application is the last ingredient in the consideration of ethics in AI. Strong ethical efforts with algorithms and data sets of the kind discussed above are not really needed in some applications. Entertainment and advertising, as already noted, can be fairly error tolerant. It is up to the user or customer. But there are applications that require much more ethical scrutiny: national security, of course (and especially the use of force); law enforcement; health care; autonomous vehicles of all kinds; fairness in credit, housing, and employment; and at least some parts of elections and political life. An "in-between" category might be commerce and some parts of finance, where the risk of error is mostly economic rather than moral and can be priced in. And even seemingly innocent applications in the consumer Internet can turn in dark directions when their true mission is deceit, manipulation, privacy violation, or their enablement.

The reason the techniques for scrutiny of AI algorithms and data sets described above are important is that the complexity and relative newness of AI can conceal ethical problems from even ethical users of technology. In this respect, AI is no different from other new technologies: they always create new capabilities that must be situated in a framework of right and wrong that itself changes slowly, or arguably not at all. Nuclear weapons, for example, certainly created new capabilities of mass destruction, but the moral principles of just war, proportionality, and discrimination still applied to them.

I believe that this discussion of AI algorithms and big data sets demonstrates that AI is not impenetrable. It is possible to locate right and wrong even in AI's amazing complexity. It is *not* possible to claim that the technology itself makes moral use indefinable. It even follows that occasional tragic outcomes are defensible if these techniques have been used with care. What is indefensible is application of AI to inherently immoral purposes, or deployment without the technical efforts described above.

I cannot rule out that someday AI might create truly new ethical puzzles for humanity. Such dilemmas would have to evade both deeply informed expert scrutiny of the technology and extant moral principles. While the popular press sometimes alleges that AI has created qualitatively new ethical quandaries, no such cases have been found to date.

There is, therefore, a variety of engineering approaches to building ethics into AI. The technical judgment that AI and big data, despite their seemingly ineffable complexity, do not defy moral examination is good news for U.S. national defense. As Secretary of Defense, I made no apology for the fact that America takes its values to the battlefield. But I also made a large number of changes in the DOD's structure and practices to connect the Pentagon more closely to the tech sector. When I took my first job in the Pentagon in 1980, most new technology, including AI, originated in America, and most of that under government (largely under DOD) sponsorship. Today, the tech base is commercial and global. For this new era, the Pentagon therefore needs to build new and different kinds of bridges to the tech sector. Accordingly, as Secretary of Defense, I founded the Defense Digital Service to bring young techies directly into the Pentagon, placed Defense Innovation Unit outposts in the nation's tech hubs, and convened a Defense Innovation Board chaired by Google's Eric Schmidt.

But in the same role I also authorized raids, hostage rescues, counterterrorist operations, ongoing combat operations in Afghanistan, the major campaign to destroy the Islamic State, and a host of war plans devised for China, Russia, Iran, and North Korea, all of them requiring grave moral judgments and all of them using the newest technology the Pentagon had. It is important that leaders be able to situate important moral principles in dramatically new technological settings, rather than being bamboozled into thinking they do not apply.

The list of exploding tech fields is long. It encompasses all forms of intelligence collection and electronic warfare, cyber warfare, robotic vehicles, ubiquitous presence via space, IoT, global WiFi and LiFi, bioengineering and biodefense, all sorts of new engineered materials, undersea warfare, microsatellites, human performance enhancement, various quantum applications, directed-energy weapons, and hypersonic vehicles. To make room for these innovations in the defense budget, familiar military capital stock like manned armored vehicles, many surface ships and large-mass satellites, manned aircraft, and even certain infantry subspecialties will gradually be phased out. The only field of warfare in which changes are not anticipated is nuclear weapons. Without exception, each of the new technologies is being developed and tested using AI. For example, new materials development rests on quantum mechanical equations of multi-atom geometries that are easy to write down but intractable to solve in closed form: AI-enabled computer calculations are the only way these new materials – with fantastic weight, strength, thermal, electronic, and other properties – can be engineered. A U.S. military unmatched in its use of AI is therefore not only essential, but also key to all kinds of military innovation.

Another question I was frequently asked as Secretary of Defense is whether there will be two Internets, one U.S.-led and one China-led. There will, indeed, surely be two tech ecosystems. That is not a choice the United States can make; Xi Jinping has announced China's intention to make it so. Moreover, in geopolitical terms, China's development has not taken the path that Americans and their allies had naively hoped for as recently as a few years ago. China has not embraced values of universal valence, as America does, at least on its best days, but instead embraces values that are distinctly and exclusively ethnocentrically Chinese. Thus, the United States and China have become locked in a titanic geostrategic struggle incomparably more complex than that between the United States and Soviet Union during the Cold War. The two Cold War opponents did not trade with each other in high-tech goods. The United States and China do.

It is essential to any U.S. Secretary of Defense that America continue to be unsurpassed in all the emerging fields of technology, including, of course, in AI. Prevailing in the competition will require a new geostrategic playbook for competition with China with chapters on defense, offense, and new alliances. Defense encompasses carefully tailored restrictions on critical sensitive technology that could make its way into China. Far more important than tech defense to limit China is tech offense to improve America: robust federal research and development funding and an overall innovative climate – encompassing regulation, education and immigration, capital markets, and so on – that is maximally simulative of superiority in AI and other fields (all, as noted, enabled by AI). Finally, recalling that China makes up but one-half of Asia and one-fifth of the world, it is essential that the U.S.-led tech ecosystem embrace most of the rest of humanity. It is unlikely that China or other potential military opponents of the United States will respect the same moral scruples that the United States applies to itself. But this essay suggests that the United States will not be disadvantaged in such an asymmetrical competition since good engineering design can accommodate both high performance and good ethics. Assuming the United States retains its historic values and does not forget to apply them to AI and other new technologies in the manner described here, the result will be a peaceful and progressive world for most.

AUTHOR'S NOTE

I am grateful to D. J. Patil, former U.S. Chief Data Scientist, for helpful comments on this essay.

ABOUT THE AUTHOR

Ash Carter, a Fellow of the American Academy since 1992, is Director of the Belfer Center for Science and International Affairs and Belfer Professor of Technology and Global Affairs at Harvard Kennedy School, and a member of the President's Council of Advisors on Science and Technology. He served as U.S. Secretary of Defense from 2015–2017. He is the author of *Inside the Five-Sided Box: Lessons from a Lifetime of Leadership in the Pentagon* (2019) and *Preventive Defense: A New Security Strategy for America* (with William J. Perry, 1997) and editor of *Keeping the Edge: Managing Defense for the Future* (with John P. White, 2001).

ENDNOTES

[1] The term AI has been around a long time and, for our purposes, means all kinds of advanced techniques: machine learning, neural networks, deep learning, and even just "big data." It does not make the distinction of "artificial general intelligence" (AGI) since the definition and meaning of AGI are not precise, and its "singularity" date–when AI matches or surpasses human intelligence–is elusive. The real singularity in the existence of technology will be when we can achieve human immortality, either digital or biological. "Immortality" might even happen before AGI.

[2] What about "full disclosure," "opt in/opt out," "anonymity," "it is impossible with such complicated systems"? All these are much more dubious, as we shall see.

[3] U.S. Department of Defense, "Autonomy in Weapon Systems," Directive No. 3000.09, November 21, 2012.