"I dreamed about a human being." (Collage By Fran Simó)

Project Governance for Defense Applications of Artificial Intelligence

An Ethics-Based Approach

By Brian T. Molloy

The recent Department of Defense (DOD) Artificial Intelligence (AI) strategy calls for the Joint Artificial Intelligence Center (JAIC) to take the lead in "AI ethics and safety."[1] In line with this directive, the JAIC and the individual services must develop a coherent ethical review process to identify and mitigate potential ethical risks during project development. To date, the U.S. Defense Innovation Board (DIB) has handled much of the DOD emphasis on AI ethics culminating in the publication of its AI principles report.[2] Although it provides guideposts, it does not necessarily generate actionable controls to limit ethical risk on individual projects. The challenge for the department is to generate a project governance architecture that adequately addresses these ethical risks while also reaping the considerable benefits of AI. This article provides recommendations for implementing project governance controls based on an ethical framework while providing tailorable solutions to tightly control those projects with high ethical risk and speeding the implementation of those with low risk. In this way, a tiered approach to project governance will allow the Department to more closely balance the ethical challenges with the need for efficiency in the development of this technology.

Not all AI projects carry the same ethical risk, yet DOD currently lacks a formalized process to delineate and separate projects by ethical risk or consequences or both. Currently, the Department draws a distinction between Lethal Autonomous or semi-autonomous Weapons Systems (LAWS)[3] and those that are not autonomous weapons systems, yet there is more distinction that needs to be made in order to adequately identify the risks associated with the technology. In tiering out risk categories, the Department can focus resources to review and limit ethical risk on those projects most likely to cause ethical dilemmas while accelerating those identified as low consequence. It should be noted that all projects pose their own unique ethical concerns and that there is no one-size-fits-all policy that can be applied to limit all potential ethical challenges across the broad array of projects being pursued. To address such a fundamentally important issue, this article proposes a business process that can identify ethical risks and then mitigate them appropriately, according to their relative risk.

The Defense Innovation Board AI principles report detailed 12 specific recommendations for DOD to focus on to manage ethics in AI.[4] This article focuses on two of these recommendations and proposes

MAJ Brian Molloy is the Deputy Director of Engineering at Joint Task Force - Bravo in Soto Cano Air Base, Honduras. He is a graduate of the U.S. Naval War College where he studied the Ethics of Emerging Military Technology and other defense issues.

project governance solutions to address these challenges and translate them into concrete project governance controls. First, in line with the DIB's recommendation to create a risk-based management methodology (DIB recommendation #9), this article proposes using risk-based screening criteria to separate and tier projects based on their ethical risk. This approach allows for more stringent controls on projects of high risk, while speeding through projects of low risk. Second, the article builds on the risk management tiering framework and uses this framework to provide recommendations on AI reliability benchmarks (DIB recommendation #7).

AI as a Unique System Enabler

Artificial Intelligence, as an enabler to weapons systems, is unique in its ethical concerns and considerations and warrants a new screening approach outside of those in normal acquisition channels. Unlike other weapon systems, in AI/Machine Learning (ML) projects the end use and development of the product are more closely linked. That is to say, when using AI in a weapon system, the developer will, by design, make some choices that under conventional applications would be left to the end-user. These choices are not necessarily self-evident but are emergent based on the decisions made by developers about the boundaries and rules developed within a particular algorithm. Much has been discussed regarding potential ethical issues with ceding decision-making to AI algorithms.[5,6,7] The question becomes, can we control those decisions and bound them appropriately so that we can control the ultimate end use of the system?

Traditionally, ethical controls on technology were inserted through policy constraints on end-use. However, the relationship between developer and end-user is shifting the ethical burden backwards towards the developers. This shift necessitates a new approach to managing ethical issues. It is simply not sufficient to place policy constraints on end-use. In order to adequately mitigate ethical risks with this technology, policy controls to adjudicate ethical challenges must be applied at the outset, during the design phase, and then continued during development.

This relationship between end-user and developer will be further strained by the movement from narrow AI towards more complex adaptive systems. In the traditional designer-user relationship, the design engineers allowed themselves a certain level of plausible deniability as to the intent of the end product. In effect, the engineers could pass off the ethical dilemmas to the end-users and force them to make the hard decisions. As systems become more automated, however, it will force engineers and therefore policymakers to be more upfront with the potential ethical challenges of end-use. This problem emerges from the fact that the actions of the system will be bounded by the parameters of the design engineers. In simple mechanical systems, all decisions regarding use are made by a human operator, thereby all moral decisions regarding use or non-use are pushed to the user based on context and surroundings. In highly automated systems, however, those decisions must be made by the engineers on the front end. Therefore, during the development of each system, a program of identifying risks and consequences must be developed and then implemented through both internal controls in the algorithm and external controls through policy constraints.

AI-enabled systems must be viewed through the lens of a moral agent. That is, a system that on the one hand, does "not necessarily exhibit…free will, mental states or responsibility," but on the other hand is an entity that performs actions.[8] It is these *actions*, that have ethical ramifications. The moral decisions are not made by the machine, they are made by the design engineers, and the machine is merely the agent that carries out the action expected of it. Therefore, it is here, in the development stage, that the focus of project controls and

U.S. drone attack on the convoy of the Iranian general Qassem Soleimani, 3d render. Baghdad airport, Iraq.

project governance must lie in order to effectively manage ethical risks. The most effective way to achieve ethical behavior by a moral or ethical agent would be to ensure that the outputs of the machine are constrained to avoid unethical outcomes. This could be accomplished by creating the boundary states that implicitly support the ethical behavior of the machine by not allowing the system to conduct actions that are outside of the ethical framework.[9]

An AI Ethical Risk Management Methodology

From a policy perspective, the focus of ethically applying artificial intelligence to weapon systems needs to focus on defining the boundaries for the given technologies. This article relies heavily on the utilitarian approach to ethical issues, or the view that the morally correct action is one that produces the greatest good. There is a practical reason for this. The utilitarian approach focuses on weighing risks with consequences and tends to be the approach that is most easily quantified and measured.[10] Using this approach allows program managers and policymakers the ability to make rational decisions regarding the potential risks of the technology and to make informed mitigation decisions. It is important to note that this does not foreclose the use of other applicable ethical lenses, yet it provides a clear way ahead for providing policy guidance to the development of these technologies.

The two major recommendations for consideration in risk management are discussed below. First, project acceptance criteria must be adopted in order

to initially identify the initial ethical risk and determine boundary conditions for development. Second, projects cannot be evaluated in a one-size-fits-all approach; a consequence-based project tiering must be developed which separates projects based on potential consequences and applies additional controls to those of higher consequence while allowing those of lower consequence to be moved through more quickly.

Tiering of Projects Based on Ethical Risk

The process of tiering projects for ethical risk must begin with an initial screen for ethical issues in development. Applying a utilitarian perspective, if the project's expected benefits outweigh the potential risks within the proposed boundaries, the project then may be continued for development. If the project fails to meet this test, the Department can choose to limit the boundaries of the project to a more tightly controlled problem set until the ethical balance is achieved or until the project is deemed to be irreconcilably unbalanced and discarded. Importantly the output of this initial screen should be codified in an official document such as an "Ethical Issues Report" which would determine initial bounds for development, and this report would then be updated during project execution with additional controls based on more in-depth analysis explained below.

This initial screen of projects is likely already occurring in an informal fashion, yet a formalized procedure would force the Department to codify and document guidance to program managers within the services and to continue in an institutionalized fashion the ongoing identification and mitigation of emergent risks throughout the lifecycle of projects.

Only once the risks are identified can controls be applied within the identified boundary states in order to ensure that ethical risks are effectively managed. Each project must be tiered out based on its consequence level, then scored against its potential risks. Because some risks are more relevant than others based on project consequences, a risk relevancy matrix has been developed to assist with screening project risks. The matrix presented below can be used to ensure that ethical risk management strategies are being applied appropriately based on the type of project associated.

Table 1: Risk relevancy matrix

	Consequence Tiering Level	Ethical Risk Relevancy Matrix				
Tier 1	Lethal Autonomous or Semi-Autonomous Weapons System	High	High	High	High	Low
Tier 2	Targeting information Systems	Moderate	Moderate	Moderate	High	Low
	Safety-Critical Systems	High	High	Moderate	High	Low
Tier 3	Privacy concern systems	Moderate	Moderate	Moderate	Moderate	High
	Business Process systems	Low	Moderate	Low	Low	Low
		Technical Safety	Malicious-Unintended Use case	Algorithmic Bias in data set	Algorithm training Risk	Excessive collection risk

The recommended consequence categories to be developed are: lethal autonomous or semi-autonomous weapons systems, targeting information systems, safety-critical systems, privacy concern systems, and business process systems. Those categories are then tiered to determine additional controls on projects. Tier one projects should already be identified and by Department policy are to be managed in accordance with DODD 3000.09 Autonomy in Weapon Systems.[11] Tier two projects include all projects that result in targeting information and projects that have a significant safety risk. These projects must be screened for ethical risks based on the risk categories in the table and outlined below. The risks should then be formalized and mitigated through a formal risk management procedure overseen by the AI Ethics Committee.

The recommended risk categories to be scored against are technical safety risks, malicious/unintended use case risks, algorithmic bias in data set, algorithm training risks, and excessive collection risks.[12]

1. Technical Safety: The first question for any application is whether it works as intended over time and in various expected applications. The question of the reliability of the system in context is a significant issue in AI systems. This technical safety risk is especially acute in that in a contextually sub-optimal state a system may perform an unexpected action resulting in an unethical consequence. This problem is generally mitigated through a rigorous test and evaluation process; however, for AI/ML or other complex adaptive systems, this process is challenged as discussed below. These technical safety/reliability risks pose a significant ethical concern, in that the system can only be ethically employed if its output is sufficiently known by the operator. The limitations of this technology must be well communicated to the operator in advance of the decision to employ it. Unreliable systems pose a challenge to this dynamic in that the operator may believe it to be operating normally when it is not.
2. Malicious/Unintended Use Case Risks:[13] The second risk to be analyzed is the unintended use case risk. This risk applies to a properly functioning system that is used in a way outside of the expected or approved usage. In this case, an ethically responsible application could be co-opted by end-users for potentially unethical consequences. A detailed review of possible use cases should be conducted to identify and mitigate the possible unintended uses.
3. Algorithmic Bias in Data Set: One challenging aspect of neural networks is that the data that is used to generate the outputs are generally created and curated by humans who harbor inherent biases. These data sets by their very nature have the possibility to produce unintended results. In this case, it refers to the fact that the algorithm will reflect the implicit values of the person who developed it. This is the one ethicists have focused on the most. These biases are then broken down into three subcategories or more depending on the author; pre-existing bias, technical bias, and emergent bias.[14] These biases have been in place in software engineering well prior to the advent of AI but remain significant in the use of AI.
4. Algorithm Training Risks: One major risk in this category for military applications lies in insufficient datasets to train ML algorithms on. In the case of military applications, there are simply not the number of examples that would provide the necessary context for an AI algorithm to operate in varying environments. The mitigation for this risk has primarily been to create synthetic training environments for the algorithms to operate in. The challenge with this approach is that the synthetic environment will likely not be an exact match for the

operating environment out in the world. This mismatch has the potential for the algorithm to operate outside the bounds intended.

5. Excessive Collection Risk: Algorithms that autonomously collect and analyze data have the potential to excessively collect data beyond the scope of the initial application. This excessive collection has the potential to cause privacy or even legal challenges in the use of the technology. This risk is especially prevalent in the use of AI in the cyber domain, where the data on networks is not well defined in terms of ownership or nationality.[15]

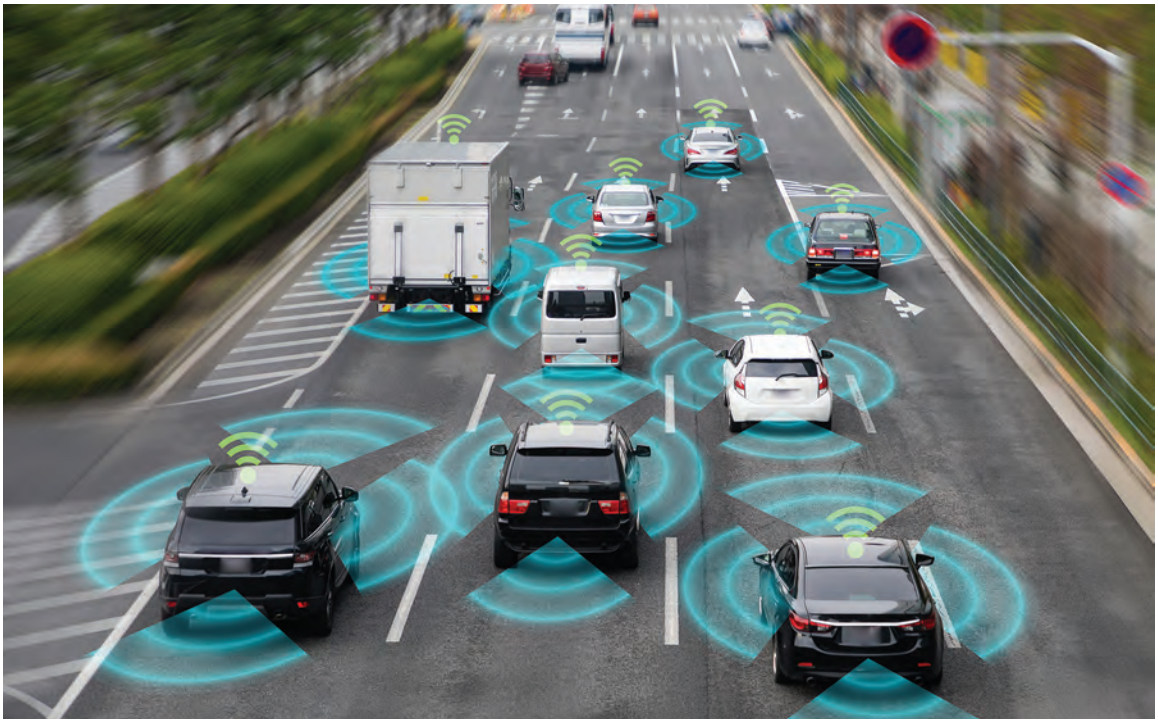
Reliability Benchmarks for Defense AI Applications

Reliability benchmarks remain one of the major unanswered questions for the development of AI-enabled systems within the DOD. AI has the potential to revolutionize the way the Department does business, but as with each new technology, there is risk associated with the adoption and widescale use of the new technology. By codifying hard reliability benchmarks, the Department can formalize the risk acceptance for developers. Likewise, defining reliability will give end-users the opportunity to understand the limits of their respective systems with greater fidelity. With this understanding, the DIB AI report recommended developing AI performance benchmarks relative to human performance.[16] The approach of tying reliability to human performance is not new; the same approach has been taken in many other industries, most notably the self-driving car industry. The DOD, however, has unique challenges that will compound the difficulty of achieving these same standards for benchmarking. For DOD applications, simply addressing whether an AI-enabled system performs better than a human analog is an insufficient approach to manage the risks associated with AI-enabled systems adequately.

Again, a tailored approach should be taken in order to manage the risks appropriately based on the risk for each application. This article proposes the use of three separate benchmarks, aligned against the project consequence tiering criteria introduced previously to more closely align the performance requirements with the potential for unintended consequences. It should also be noted that the research into AI reliability benchmarking is extremely new, and therefore there is a dearth of published industry standards or academic research on which to rely. The self-driving car industry appears to be the furthest along in this effort, but even here, many different approaches are being adopted with no single standard accepted as the norm. Some standards, such as ISO 26262, have developed highly strict standards that state that a car can only make 10 mistakes for each 1 billion hours of operation while humans are expected to make 10,000 mistakes in the same period of time.[17] Yet even this ISO standard has not been widely adopted. In this environment, applying policy ecommendations remains a challenge yet is imperative to ensure the continued viability of this technology.

Technical Challenges in Reliability Benchmarking

The unique environment that the DOD operates in compounds the problem of applying appropriate reliability benchmarks like those in other industries. The DOD environments for which AI-enabled systems are being developed are in many instances high-consequence while at the same time low frequency. The high-consequence nature of the systems will require extremely high reliability, while the low frequency of such events creates a data deficit challenge. This deficit makes it difficult to adequately train algorithms to match, or exceed, human performance. For a moment, let us consider the self-driving car industry as an analog for a high-consequence complex adaptive system. For this industry, human performance is relatively well known, and vehicle



Sensing system and wireless communication network of vehicle. (Metamorworks)

fatality data is readily available. In 2016, for example, there were 1.16 fatalities for every 100 million miles driven. To adequately field test an AI-enabled autonomous driving system to reach 95 percent reliability in comparison to human performance, the vehicle would need to be tested for 255 million miles with no accidents.[18] This standard is extremely difficult to achieve for self-driving cars even in such a data-rich environment. The self-driving car industry is one that is high-consequence but also high frequency. Yet even in this industry, novel approaches are being made in order to ensure reliability that approaches or exceeds human performance. Even in such data-rich environments challenges exist in generating data to match human performance, and the industry has begun to rely on a significant amount of synthetic data, or data that is created in a simulated environment, in addition to real-world test miles. In this industry, the standard approach uses a vast amount of raw data in order to ensure reliability.

Contrast this to a combat environment, where the accumulation of data is incredibly difficult. The environment is extremely data-poor, resulting in a significant challenge for field testing to determine system reliability.[19] Self-driving cars may see hundreds of thousands of examples of stop signs in all manner of environments, orientations, partial obscurations, and defacements during field testing. Combat vehicles will not have such data available. Consider for a moment a Russian T-90 Armada tank. How many examples would it take to achieve human parity with the identification and classification of such a threat enemy combat system? Now let's consider the number of cases where field testing data can be developed. The number of actual meeting engagements with Russian tanks as example data is infinitesimally small compared with the number of stop signs. Many novel training approaches are being developed to address this problem, including synthetic

or simulated training environments, but these approaches engender their own risks.[20]

Applying a human condition as a benchmark is likewise equally challenging. Indeed, for each system, the failure rate of humans must first be measured and then translated into requirements for the machine to be benchmarked against. The measurement of such failure rates could be incredibly difficult and, in some cases, misleading. In many cases, this type of analysis falls victim to what is called the "human filter pitfall."[21] In these cases, using human failure rates as a benchmark against machine performance can be challenging because humans and machines perceive their environment in different ways. Machines and humans operating in the same environment may have wildly different failure rates based on their respective limitations. In many cases, the machine may perform exceptionally in areas that humans routinely fail at and thus meet the reliability benchmark, yet routinely fail at other common tasks that humans do not find difficult.[22]

Additionally, the ability to accurately measure human performance, or even to determine what parameters human performance should be measured against, is difficult. Consider again the same ML algorithm designed to identify T-90 tanks for a ground combat system. How do we determine *human* reliability benchmarks for this relatively narrow task? Since the current physical training environment is data-poor, that is there are very few actual T-90 tanks for soldiers to look at in person, the average soldier is essentially trained on flashcards of T-90 tank photos. We could, therefore, base the reliability standard on the average soldier's ability to accurately identify T-90 tanks in these photos in various environments. Yet when the soldier, or the ML algorithm, encounters this in the field, the reliability changes dramatically. A soldier under the stress of combat will have remarkably different reliability in this task.[23] Indeed, the soldier in the rush of combat may not see the tank at all, an error of omission. Or they may commit errors of commission, that is, misidentify friendly tanks as enemy or enemy tanks as friendly. In other cases, the soldier may identify the tank, but choose not to engage due to some other reason. Perhaps the proximity of non-combatants, perhaps there were other nearby targets or other indications that the tank was not a threat. In these cases, measuring human reliability becomes increasingly challenging. Determining what metric to use as a *human* performance standard must be addressed along with the reliability standards for machines.

Social Acceptability of Reliability Benchmarks

The public's willingness to accept technological change further exacerbates the policy risk that mistakes may impose for the use of this technology. Indeed, any discussion of reliability ultimately can be distilled into a discussion of risk and risk tolerances. Defining a reliability metric that constrains the use of technology in only those cases where it will always outperform a human is one approach to limit risk and manage the risk tolerances of the public. Yet the risk tolerances of the public are not entirely rational. In many cases, the public appears to have a lower risk tolerance for technologies that retain certain characteristics. In some cases, this manifests as technologies that generate visceral emotions.[24] In others, the effect is seen where mistakes cannot be easily explained.[25] In all of these cases, the perceived risk tends to skew much higher than the actual risk.[26] Several heuristics account for the risk perceptions being skewed that are particularly relevant to AI technology. This effect is even more pronounced with new or novel technologies that are not easily explained; this effect has been labeled as "new-risk."[27] This new-risk phenomenon is particularly relevant for AI-enabled systems. A significant perception exists that AI is such a new and untested technology that it simply cannot be trusted, especially in applications with high-consequence. This fear tends to outweigh opinions on the suitability for the application even when presented with

evidence that the technology will ultimately save lives.

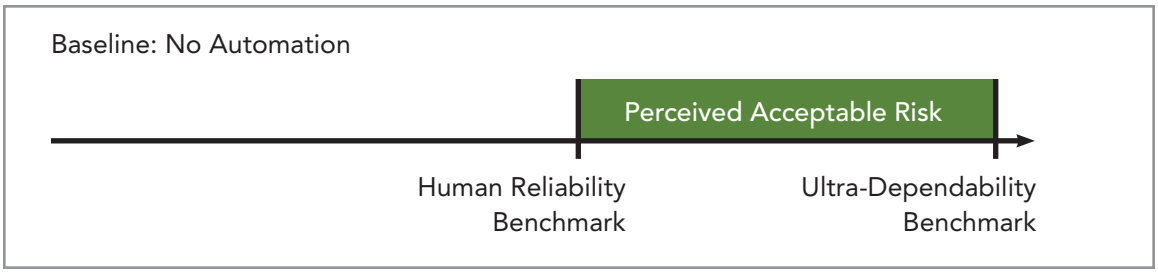
The other relevant heuristic appears to be what is known as "Unnatural or Immoral risk." In other words, technologies that are deemed to have high ethical risk are viewed as riskier than those of other types of risk. This immoral risk phenomenon originated in the nuclear industry,[28] but can easily be extrapolated to other novel military technologies that were ultimately deemed immoral. Many have said that military applications of AI will also fall into this same category. Global movements against autonomous weapons systems are prime examples of the immoral risk perception that accompanies AI in the military sector.[29] It should be noted that the resistance to autonomous weapons systems has followed the reasoning used for the banning of other novel technologies deemed immoral, namely chemical weapons, biological weapons, and landmines. In the case of AI, the question of responsibility and even whether it is morally acceptable or even a violation of human dignity to be killed by a robot are ethical questions that are being hotly debated.[30] Both the newness and the perceived immorality combine to form a *trust gap* that must be overcome in DOD policy.

For most policy documents, the utilitarian approach to develop control mechanisms is generally appropriate. Yet in this case, this approach would lead to a policy mismatch with popular opinion. The utilitarian approach to this problem would be to say that any AI-enabled tech that can outperform a human analog should be allowed to be fielded. In effect, the benefit of the increased performance would outweigh the risks. Yet the perception of this technology does not necessarily follow purely utilitarian perspectives. This mismatch sets up a trust gap that must be overcome in order to achieve public support. The deployment of this technology without public support would put at risk the continued use of the technology and could stymie research and development efforts with wide ranging consequences.

To explain this dynamic, it is important to define some terminology. The first concept is the human reliability benchmark, which represents human analog performance for a tightly controlled task as compared to an AI-enabled system performing the same task. The second concept is the ultra-reliability benchmark. This benchmark, a term borrowed from the airline industry, is a no-fail benchmark since any failure would be considered unacceptable. Essentially, this ultra-reliability benchmark connotes a hypothetical standard where no failures should occur under any circumstances.

The control set for this analysis would be an entirely human-controlled system. This is essentially the system the DOD has operated under since its inception. Under this construct, the population, and DOD policy, understand the limitations of soldiers under the stress of combat and allow for mistakes to be made. The public accepts that human performance will never reach the no-fail ultra-reliability standard. The space between these two systems is defined as the perceived acceptable risk. This acceptable risk can be extrapolated out into perceived *policy* risk.

Figure 1: Baseline Risk Acceptability



However, as the combat environment has continued to become more complex, and weapons systems have become more automated, a new layer of reliability threshold emerges. Here we may define the emergent threshold as an acceptable *machine* reliability benchmark. The space between the human reliability benchmark and the machine reliability benchmark illustrates a trust gap. This gap is somewhat counterintuitive but can be explained by the lack of risk tolerance by the public for "new-risk," or in the case of military technologies, "immoral-risk." In these cases, some mistakes that were acceptable for a human operator are not acceptable for a machine performing the same task. It can be expected, therefore that for technologies with relatively low consequences, the immoral-risk factor will be lower, resulting in a smaller trust gap. This approach explains that benchmarking machine performance to simply match human performance may not be adequate to overcome the perceived acceptable risk of emerging technologies.

An ethical framework analysis can partially explain the emergence of this trust gap. Under a utilitarian model, this trust gap would cease to exist; it should not matter whether a human or a machine was performing the task if the only thing that matters is the result. If a machine with high consequences outperforms a human operator, then by a purely utilitarian logic it would be immoral not to field the system. Yet we do not live in a purely utilitarian world. As noted above, various heuristics are at play. One of the most powerful is the idea of immoral risk. This risk perception relies not on utilitarianism but on virtue ethics.[31] It is understood that humans have virtues; whether machines can have virtues remains an open question. Here, the question becomes whether the public will accept a mistake made by a potentially unvirtuous machine less frequently or by an ostensibly virtuous human more often. For these reasons, it can also be expected that as automation increases, the trust gap will also widen. This phenomenon is largely due to the idea that with a small amount of automation, humans remain largely in control. Yet as the level of automation increases, or the consequences of the task being automated increases, people are more likely to be dubious of the ability of the machine to act as a moral agent.

This scenario becomes even more pronounced when the technology has high consequences. This dynamic can be seen playing out in real-time once again in the self-driving car industry. In this case, the public's acceptance of perceived acceptable risk is exceedingly small. The self-driving car industry, like many DOD programs, is viewed by the public as a technology that is highly automated and high-consequence. In this scenario, both the new-risk and the immoral risk weigh heavily on public opinion and push the acceptable reliability thresholds to reach far beyond human performance. In this case, here defined as automation with high consequences, the trust gap is large and must be overcome by setting a reliability benchmark much higher than human performance. This trust gap is consistent with the

Figure 2: Risk Acceptability in Automation with Low Consequences

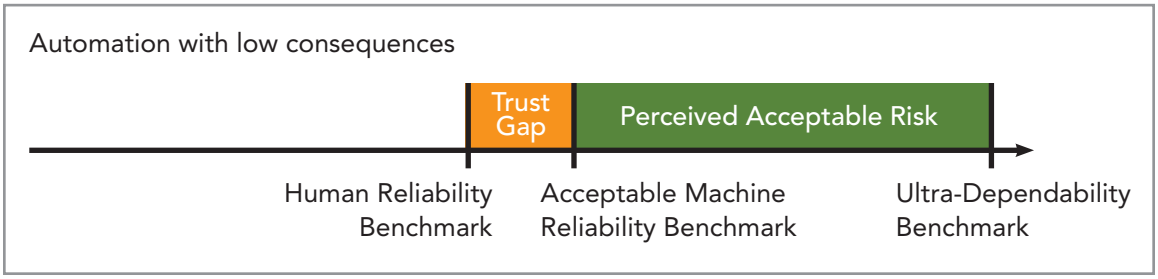
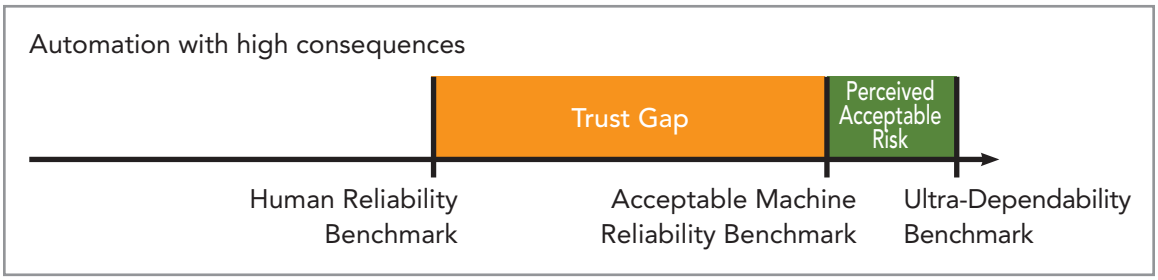


Figure 3: Automation with High-consequences



ethical risk relevancy matrix outlined above. It should be noted that those projects that were identified as having higher ethical risk would also fall victim to the trust gap detailed below.

The DOD must develop reliability benchmarks with these factors in mind. The same consequence tiering levels are recommended in order to adequately allow rapid development of those projects of lower risk while more tightly controlling those of higher risk. In setting these benchmarks, the Department must balance the potentially profound implications that high-consequence mistakes may have on the overall use of the technology with the ability to enjoy the benefits that the technology promises. It has been well documented that the early deployment of this type of technology can lead to exponential increases in overall safety. A recent study by RAND conducted a detailed analysis of this very question for self-driving cars and found that the deployment of self-driving cars at a 10 percent improvement over the human condition would result in significant savings. The report states that "more lives are cumulatively saved under the less stringent … policy than the more stringent … policies in nearly all conditions." The report further compared the benefits of early adoption at 10 percent improvement over human reliability to a 75 percent and 90 percent improvement and found that an early adoption strategy had the result of saving, "tens of thousands to hundreds of thousands of lives."[32] Yet this report also argues that this approach is a purely utilitarian model and cautions against relying on it alone. For the same reasons outlined above, the report recommends policymakers determine a middle ground whereby the policy is acceptable to the public, while still allowing for innovation and rapid adoption.

This article proposes the following reliability benchmarks: Tier 1 projects represent the greatest trust gap that must be overcome in order to enjoy public approval and therefore, must be the most tightly controlled. Therefore, a significant improvement over the human condition is recommended before allowing full fielding. Tier 2 projects engender much less consequence, and therefore would have a lower trust gap, and can, therefore, be less tightly controlled. For these projects, consistent with a rapid advancement model, a 10 percent improvement over human condition should be used. Finally, for Tier three projects where the consequence of failure is low and where there is little to no trust gap, it is not recommended to tie reliability metrics to human performance.

Tier 1 – These project categories have huge consequences as well as high ethical risks. The unique characteristics inherent in autonomous weapons systems mean that a purely utilitarian approach with a rapid adoption model must be avoided. As discussed above, for these projects, a very large trust gap must be overcome before any mistakes are deemed acceptable by the public. Thus, for example, a major backlash against the use of lethal autonomous weapons is likely even for mistakes made that would be easily explained as

Table 2: Benchmarking Projects by Tier

	Consequence Tiering Level	Proposed Reliability Benchmarks
Tier 1	Lethal Autonomous or Semi-Autonomous Weapons System	50%-75% improvement over human performance
Tier 2	Targeting information Systems	10% improvement over human performance
	Safety-Critical Systems	
Tier 3	Privacy concern systems	No human-based reliability benchmark
	Business Process systems	

human error in legacy systems. The potential for this backlash could grind all AI-enabled system development to a halt—resulting in the potential to lose many of the benefits of this technology.

For this tier of project, a 50-75 percent improvement over human performance is recommended. While it is understood that an early adoption methodology that matches, or just slightly improves on, human performance would result in more rapid development, policymakers must work to find an acceptable middle ground that shows a marked improvement over human performance. The early adoption models have been shown to be effective for much more narrowly defined problem sets with a smaller trust gap than can be expected for a lethal autonomous system. At the same time, the policy cannot constrain the technology to a point where the perfect becomes the enemy of the good. As was demonstrated by the RAND report on autonomous vehicles, waiting until a 90 percent improvement over the human benchmark provided marginal gains over a more modest model with a huge tradeoff in time.[33] Policymakers must find a middle ground, and a 50-75 percent increase in performance over human benchmarks will allow the DOD to cover the trust gap while still reaping the long-term benefits of the technology.

Tier 2 – A 10 percent improvement over human performance is recommended for projects with safety critical systems or those that provide targeting data. For these systems, utilitarianism wins out. The benefits gained by early adoption are more important for the DOD than the risks engendered by the potential for mistakes. Logic argues that performance must equal or exceed equivalency to human performance for the technology to make sense to field. Yet this technology still falls victim to the new-risk phenomenon and holds a small trust gap that must be overcome for both the users of the technology and the public. In order to overcome this gap while still retaining the benefits of an early adoption strategy, a modest increase over human performance is prudent.

Tier 3 – For those tier three projects which have low consequence, or those with little ethical risk, it is not recommended to tie reliability metrics to human performance. In these cases, the benefits of early adoption and continued development far outweigh the risk of mistakes. Here, the decision becomes one of functionality and suitability to the task rather than the relative comparisons to human performance in the task. Because of this, it is not recommended to place any restrictions on these applications.

Conclusion

The Department has made significant strides towards the adoption of Artificial Intelligence into critical applications across the force. These technologies have the potential to be game-changers in the way the U.S. military fights its future wars;[34] however, by their very nature, they engender significant ethical challenges. At this early stage of development, the Department has the opportunity to achieve its stated goal of becoming the leader in ethics for military applications. The institutional risks of not getting ahead of the ethical challenges are stark. If the Department runs afoul of industry ethical frameworks, it risks alienating industry, forcing broad policy restrictions from political leaders, or legal challenges to its implementation. Each of these challenges has the potential to grind AI incorporation into the force to a halt. It is, therefore, critical that the Department gains and maintains the strategic messaging that it is pursuing this technology in an ethical fashion and that its use will benefit the United States.

It is time to move past broad ideas of ethical AI in principle and translate these principles into actionable controls that will keep the advancement of this technology inside the bounds of ethical behavior. DOD must create policies which govern the ethical development of this technology. By using a risk management methodology, as detailed above, it will allow the Department to tightly control those projects of high risk while allowing low risk projects to speed through the system. In doing so, the Department can push the boundaries at the technological edge with projects of low consequence while tightly controlling those of high consequence. This risk-based framework lends itself to applying a myriad of controls on projects including the reliability benchmarks and test and evaluation policies detailed in this article. The Department, starting with the JAIC, must institutionalize ethics as part of its ongoing, routine business practices. This focus cannot be viewed as a barrier to development or a bureaucratic process that slows implementation but instead as an essential task to smoothly incorporate AI into the force. The ethical challenges with AI are not insurmountable; they do, however, need to be addressed and mitigated in a formalized fashion for the adoption of this technology to move forward. PRISM

Notes

[1] U.S. Department of Defense, *DOD Records Management Program, Summary of the Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (Washington, DC: DOD 2018).

[2] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, Office of the Secretary of Defense (Washington, DC: Defense Innovation Board, 2019).

[3] Department of Defense, *DOD Records Management Program*, Department of Defense Directive (DODD) 3000.09 (Washington, DC: DOD, 11 November 2012).

[4] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense - Supporting Document*, Office of the Secretary of Defense, (Washington, DC: Defense Innovation Board, Oct 2019), 42-44.

[5] Abigail Beal, "It's Time to Address Artificial Intelligence's Ethical Problems," *Wired*, August 24, 2018, https://www.wired.co.uk/article/artificial-intelligence-ethical-framework.

[6] Future of Life Institute, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," *Future of Life Institute*, July 28, 2015, https://futureoflife.org/open-letter-autonomous-weapons/.

[7] Brian Green, "Artificial Intelligence and Ethics," *Markkula Center for Applied Ethics*, November 3, 2017, https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics/.

[8] Gianmarco Veruggio and Fiorello Operto, "Roboethics: A Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics" in *Machine Ethics and Robot Ethics*, ed. Wendell Wallach and Peter Assaro, (New York: Routledge, 2017), 81.

[9] Ibid, 81.

[10] Lawrence M. Hinman, *Ethics: A Pluralistic Approach to Moral Theory*, 5th ed, (Boston, MA: Wadsworth Publishing 2013) 124.

[11] Department of Defense, *DOD Records Management Program*, Department of Defense Directive (DODD) 3000.09 (Washington, DC: DOD, November 11, 2012), 1.

[12] Brian Green, "Artificial Intelligence and Ethics," *Markkula Center for Applied Ethics*, November 3, 2017, https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics/

[13] Miles Brundage, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *Center for New American Security*, February 2018. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

[14] Batya Friedman and Helen Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems* 14, no. 3 (July 1996): 330 –347.

[15] Darren Shou, "The Next Big Privacy Hurdle? Teaching AI to Forget," *Wired*, June 12, 2019, https://www.wired.com/story/the-next-big-privacy-hurdle-teaching-ai-to-forget/

[16] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense - Supporting Document*, Office of the Secretary of Defense (Washington, DC: Defense Innovation Board, October 2019), 43.

[17] Manish Gupta, "Self-Driving Cars: Reliability Challenges, Solutions, and Social Adoption: Strict Reliability Standards are Critical for Human Safety and to Help Drive Social Acceptance of Nascent Self-driving Technology," *Design News*, June 22, 2018, https://www.designnews.com/electronics-test/self-driving-cars-reliability-challenges-solutions-and-social-adoption/87842508158921

[18] Philip Koopman, Aaron Kane, and Jen Black, "Credible Autonomy safety Argumentation," *Published by the Safety-Critical Systems Club*, 2019, 15. https://users.ece.cmu.edu/~koopman/pubs/Koopman19_SSS_CredibleSafetyArgumentation.pdf

[19] Robert Bond, "Artificial Intelligence for National Security Applications" (PowerPoint presentation, Massachusetts Institute of Technology, Lincoln Laboratory, Concord, MA, October 30 ,2019).

[20] Philip Koopman, Aaron Kane and Jen Black, "Credible Autonomy safety Argumentation," *Published by the Safety-Critical Systems Club*, 2019: 19. https://users.ece.cmu.edu/~koopman/pubs/Koopman19_SSS_CredibleSafetyArgumentation.pdf

[21] Ibid, 13.

[22] Ibid, 13.

[23] Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton, FL: Chapman and Hall/CRC, 2009), 29-36.

[24] Paul Slovic and Ellen Peters, "Risk Perception and Affect," *Current Directions in Psychological Science* 15, no. 6 (December 2006): 323.

[25] Berkely Dietvorst and Massey Simmons, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology* 144, no.1 (February 2015): 5

[26] Lennart Sjöberg, "Factors in Risk Perception," *Risk Analysis* 20, no. 1 (2000): 1.

[27] Ibid, 4.

[28] Ibid, 4.

[29] "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," *Future of Life Institute*, July 28, 2015, https://futureoflife.org/open-letter-autonomous-weapons/

[30] Patrick Lin, "Do Killer Robots Violate Human Rights?: When Machines are Anthropomorphized, We Risk Applying a Human Standard that Should not Apply to Mere Tools," *The Atlantic*, April 20, 2015, https://www.theatlantic.com/technology/archive/2015/04/do-killer-robots-violate-human-rights/390033/

[31] Virtue ethics is an ethical framework which is rooted in the human condition. Essentially it requires a human to have the right moral virtues, and to then be able to use intelligent judgement to conform those virtues to a given situation, and then to ensure that the proper means are used to carry out actions. For more on virtue ethics see: Alasdair Macintyre, "Virtue Ethics," in *Encyclopedia of Ethics*, ed. Lawrence C. Becker and Charlotte B. Becker, 2nd ed. (City: Routledge, 2001), 1276-1281.

[32] Nidhi Kalra and David G. Groves, *The Enemy of the Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles* (Santa Monica, CA: RAND Corporation, 2017), 25.

[33] Ibid, 19.

[34] Shawn Brimley, Ben FitzGerald and Kelley Sayler. "Game Changers: Disruptive Technology and US Defense Strategy, "in *Disruptive Defense Papers*, ed. Thomas P. Hughes and Agatha Hughes (Washington, D.C.: Center for a New American Security, 2013), 3–24.

THE FOREIGN SERVICE JOURNAL

Covering the intersection of diplomacy and defense for a century, *The Foreign Service Journal* provides an insider's perspective for foreign affairs professionals.

Published by the American Foreign Service Association.

Visit us online at **www.afsa.org/fsj**

Subscribe at **www.afsa.org/subscribe-fsj**

Now online:
The full 100-year archive of *The Foreign Service Journal*
www.afsa.org/fsj-archive