JUNE 2018

TECHNOLOGY ROULETTE

Managing Loss of Control as Many Militaries Pursue Technological Superiority

Richard Danzig

Center for a New American Security

About the Author

RICHARD DANZIG is a Senior Fellow at the Johns Hopkins University Applied Physics Laboratory. He was Secretary of the Navy in the Clinton Administration.

Acknowledgments

This paper began as a small section added to a more general study I undertook for the Intelligence Advanced Projects Activity (IARPA). That study addressed the challenges for the United States of maintaining superiority in the absorption and development of technologies relevant to national security. Jason Matheny, the Director of IARPA, supported that work and has been an invaluable proponent and commentator on the work as it has developed. I am grateful to him for his exceptional intellect, character and concern about these issues.

The Open Philanthropy Project supported broadening and deepening this part of the IARPA paper and preparing it for general publication. The Center for a New American Security undertook that publication. I am particularly indebted to Helen Toner and Claire Zabel at Open Philanthropy and to Paul Scharre and Loren DeJonge Schulman at CNAS for their contributions. I would like to thank Kara Frederick, Maura McCarthy, and Melody Cook for their role in the review, production, and design of this report.

Open Philanthropy's support included sponsorship of a workshop earlier this year in which a draft of the paper was discussed by a number of participants with special insight. I am grateful to Jeff Alstott, Jim Baker, Jack Clark, Allan Dafoe, Craig Fields, Dan Geer, Larry Gershwin, Chris Inglis, Jason Matheny, Tim Maurer, Tyson Meadors, Jim Miller, Ernest Moniz, Anne Neuberger, David Relman, Peter Singer, Helen Toner, Renee Wegrzyn, Bob Work, and Philip Zelikow for their energetic participation. Craig Fields, Dan Geer and Phil Zelikow went beyond this by providing detailed follow up comments.

In addition, Chris Beinecke, Steve Burton, Miles Brundage, Paul Gewirtz, Michael Hopmeier, Lou Pringle, Al Martin, Alex Montgomery, and Jonathan Reiber all commented on the manuscript and enriched my thinking. Peter Levin offered his usual fine blend of trenchant criticism, editorial suggestions, and encouragement. Professor Jack Goldsmith helped me by assigning the paper to his Harvard Law School class and eliciting useful comments from his students. I am grateful to all of them and to Chris Kirchhoff who helped lead this discussion when a Boston snowstorm kept me from attending. Alex Downes provided valuable insight on the political science literature relating to norms. Terry Leighton was particularly helpful in resolutely pursuing the challenging question of what caused the 1957 H1N1 global influenza pandemic. I am grateful as well to Gigi Kwik Gronvall and Tom Inglesby for their correspondence with me on this subject.

Of course, none of those mentioned have responsibility for the views expressed here or any errors that remain. I am grateful to all for their considerable help.

Cover Photo
Getty Images adapted by CNAS

TECHNOLOGY ROULETTE

Managing Loss of Control as Many Militaries Pursue Technological Superiority

02 Executive Summary

04 Introduction

06 Differentiating Between Lay Fear and Expert Fear

07 Why Focus on Military Development and Use of Technologies?

09 Causes of Loss of Control

12 Inability to Reliably and Persuasively Measure These Risks

13 Humans in the Loop

17 What Is to Be Done?

21 Conclusion

22 Appendix

23 End Notes

Executive Summary

This report recognizes the imperatives that inspire the U.S. military's pursuit of technological superiority over all potential adversaries. These pages emphasize, however, that superiority is not synonymous with security. Experience with nuclear weapons, aviation, and digital information systems should inform discussion about current efforts to control artificial intelligence (AI), synthetic biology, and autonomous systems. In this light, the most reasonable expectation is that the introduction of complex, opaque, novel, and interactive technologies will produce accidents, emergent effects, and sabotage. In sum, on a number of occasions and in a number of ways, the American national security establishment will lose control of what it creates.

A strong justification for our pursuit of technological superiority is that this superiority will enhance our deterrent power. But deterrence is a strategy for reducing attacks, not accidents; it discourages malevolence, not inadvertence. In fact, technological proliferation almost invariably closely follows technological innovation. Our risks from resulting growth in the number and complexity of interactions are amplified by the fact that proliferation places great destructive power in the hands of others whose safety priorities and standards are likely to be less ambitious and less well funded than ours.

Accordingly, progress toward our primary goal, superiority, should be expected to increase rather than reduce collateral risks of loss of control. This report contends that, unfortunately, we cannot reliably estimate the resulting risks. Worse, there are no apparent paths for eliminating them or even keeping them from increasing. The benefit of an often referenced recourse, keeping "humans in the loop" of operations involving new technologies, appears on inspection to be of little and declining benefit.

We are not, however, impotent. With more attention the American military at least can dampen the likely increase in accidents and moderate adverse consequences when they occur. Presuming that the United States will be a victim of accidents, emergent effects, and sabotage, America should improve its planning for coping with these consequences. This will involve reallocating some of the immense energy our military invests in preparing for malevolence to planning for situations arising from inadvertent actions and interactions.

The U.S. Department of Defense and intelligence agencies also must design technologies and systems not just for efficacy and efficiency, but also with more attention to attributes that can mitigate the consequences of failure and facilitate resilient recovery. The pervasive insecurity of digital information systems should be an object demonstration that it is exceedingly costly, perhaps impossible, to attempt to counter loss of control after we have become dependent on a new technology, rather than at the time of design.

Most demandingly, the United States also must work with its opponents to facilitate their control and minimize their risks of accidents. Twenty-first century technologies are global not just in their distribution, but also in their consequences. Pathogens, AI systems, computer viruses, and radiation that others may accidentally release could become as much our problem as theirs. Agreed reporting systems, shared controls, common contingency plans, norms, and treaties must be pursued as means of moderating our numerous mutual risks. The difficulty of taking these important steps should remind us that our greatest challenges are not in constructing our relationships to technologies, it is in constructing our relationships with each other.

These arguments are made to the national security community. These reflections and recommendations, however, should transcend their particulars and have implications for all discussion about control of dangerous new technologies.

“Unlike a well-defined, precise game like Russian roulette, where the risks are visible to anyone capable of multiplying and dividing by six, one does not observe the barrel of reality.”

—Nassim Nicholas Taleb

Introduction

Innovations in technology and in warfare have long, perhaps always, been intertwined. The 20th century's world wars provided object lessons of the connection. World War II especially shaped our present perspectives. All military strategists recognized then, and remember now, that the war started with Blitzkrieg, a German strategy that exploited the combustion engine, message encryption, and the radio, and concluded with the code breakers in the U.K.'s Bletchley Park and the atomic bomb, a remarkable American orchestration of science and engineering.

At the end of that conflict, the United States had immense advantages – among them an unscathed (indeed, economically revived) industrial establishment, half of the world's GDP, and a majority of the global science and technology establishment. For three quarters of a century, buttressed by a plethora of postwar institutions, a vibrant commercial sector, preeminent academic institutions, and well-funded government programs, our national security strategy extended and exploited those advantages.[1]

Innovations in technology and in warfare have long, perhaps always, been intertwined.

As the world has normalized, this triad of technological, economic, and military superiority has come under pressure. With a quarter of the world's GDP, our position remains privileged, but not nearly so privileged as previously. Indeed, China's GDP is projected to be 50 percent greater than ours by the middle of this century. In modern times the United States has never faced an opponent of comparable, not to mention greater, wealth.[2] Moreover, scientists, technological know-how, and commercial production are now globally distributed. Other nations, foreign companies, and even terrorist and criminal groups have the wherewithal to purchase, create, reverse-engineer, and steal technical products, insights, and methods. It might be said that, in this respect, they are trying to be Americans. We respond to this by redoubling our efforts to maintain technological superiority.

Much is being written and said now about why, how, and whether America can sustain its preeminence. This report does not contribute to that discussion. It does not prescribe the particulars of how we should proceed; it assumes that our national security agencies will double down on our efforts to sustain supremacy; it even presumes their success.

The focus in these pages is on a collateral, but consequential, set of risks that rise when America and, following its lead, many other countries commit to using new technologies to expand their military power. These risks include unintended consequences of complex systems, errors in analysis or operation, interactive effects between separately developed systems, and distortions introduced by sabotage.

The multinational reliance on ever-advancing technological capabilities is like loading increasing numbers of bullets into increasing numbers of revolvers held to the head of humanity. Even if no one ever wants a gun to go off, inadvertent discharges may occur. One nation's inadvertent release of a pathogen, computer virus, or autonomous AI agent may have worldwide effects. Damaging, perhaps catastrophically damaging, unto itself, an inadvertent act also may trigger other reactions. For example, a Russian, Chinese, American, Israeli, or Iranian computer exploit intended to harvest intelligence might unexpectedly disrupt, or be thought to disable, a critical part of an electric grid or financial system and provoke catastrophic retaliation.

These examples do not pretend to be predictions. They are merely illustrative. The focus in this essay is on conditions in the environment that encourage unpredictable outcomes with unmeasurable probabilities, particularly those with potentially large-scale consequences.

As Taleb says in the epigraph, we can't count the chambers in the barrels we face. But increasing the number of guns and bullets seems certainly to increase our danger. This could be described as a game of Russian roulette. But because the Russian game normally involves playing with only one bullet, this report distinguishes a new game and calls it Technology Roulette. The game is one played every day all over the planet, with increasing numbers of players, ever-greater variety in our weaponry,[3] and increasing interactions within a crowded technological space, so that a live bullet may trigger deadly exchanges among many (or all) players.

Giving this dynamic a name isn't likely to change the situation. These dangers have been recognized before,[4] particularly in analyses of American nuclear command and control.[5] They can be seen, in a different guise, in the vulnerabilities introduced by digital information technologies. A first challenge is to harvest insight about risk from these experiences and apply it to newer

technological initiatives, including notably AI, synthetic biology, and autonomous systems. American national security agencies then must recognize that our risks will rise as these technologies diffuse to other nations, that these nations may have lesser safeguards, and that their and our technologies will create risk as our systems interact. Against this backdrop, American agencies need to take initiatives to reduce risks from our own R&D programs and operations, and they need to reflect on how we might induce others to do the same. This report does these things, no doubt imperfectly, but hopefully usefully.

Most importantly, these pages convey an overarching message that is as difficult to accept, as it is – in the author's view – fundamental. National security professionals focus on the struggle for superiority, emphasizing that superiority offers the strongest security and that deterrence, empowered by superiority, is our most reliable tool. These arguments are both correct and congenial. They permit these professionals to focus where they are most capable and comfortable – on winning a competition, rather than transcending it or recognizing its dangers. But they are not sufficient. Accidents and inadvertent "emergent" effects create risks outside the framework of the competition.[6] Deterrence, though indispensable, is at best only marginally relevant to reducing the risks of accidents or unintended effects. In short, superiority is not synonymous with security.

The scientists and engineers who develop technology opportunities have their own tendencies that also often obscure the implications of our technology risks. They frequently opt to discuss dangers in contexts less compelling and complex, more abstract and less subject to the sensitivities of patriotism, than presented by military programs. Those concerned with the currently most debated technology, AI, see risks from autonomous military weapons, but they unintentionally obscure urgent present issues by asserting that the future danger substantially derives from removing the human "from the loop." This misleads in three respects: It underestimates the extent to which machine decisionmaking already affects military decisions; it overestimates the likely gain from assigning humans as decisionmakers in otherwise automated systems; and it deflects attention from the fact that aspects of the challenges that appear singular to AI are also posed by other technologies – as, for example, when biologists create new pathogens that have their own autonomy.

This report argues that our national security agencies' potential for loss of control of technologies – producing outcomes that they do not intend – is a risk not only for the future. It imperils our present. Unfortunately, the uncertainties surrounding the use and interaction of new military technologies are not subject to confident calculation or control. There are no apparent paths for eliminating our risks or even keeping them from rising. More human control over operational decisions is often not a viable, or even desirable, solution. There are, however, some technological and policy steps that can moderate our risks of loss of control. The United States should give them much more priority than at present. The hardest part of doing that is not in constructing our relationships to these technologies, it is in constructing our relationships with each other.

Superiority is not synonymous with security.

The argument and recommendations are presented in six sections and a conclusion. Part I points out that laymen commonly fear consequences from new technologies because they are unfamiliar, unnerving, and destabilizing to existing processes and power relationships. It notes, however, how those not usually distracted by these considerations, the experts and advocates for the new technologies highlighted here, are now, uncommonly, the very ones in the vanguard of emphasizing catastrophic risks inherent in these technologies. Part II specifies why, not withstanding valuable training and routinized precautionary procedures, military use of these technologies poses particularly great risks of producing inadvertent, undesired outcomes. The next part outlines the five principal paths by which these outcomes have been, and may well again be, produced. Part IV describes why these risks of loss of control cannot be calculated with even minimal confidence. Part V examines the advantages of the "human in the loop" and argues that, though this perceived safeguard yields advantages, these benefits are presently very limited and likely to be even more circumscribed in the future. Part VI argues that amidst this uncertainty we can and must do much more to control military technologies. Five actions are specified. A concluding section articulates some fundamental lessons that flow from this analysis.

Differentiating Between Lay Fear and Expert Fear

It is easy to point to clarion calls alerting us to technological apocalypse as a result of new technologies. Most of these simply express the emotions of those who are unsettled by the unknown. Fear is a companion to technological progress. New technologies commonly come to users unbidden – Trojan horses brought within the walls of established operations.[7] Competitive pressures make it impossible to refuse these advances. Once admitted, they are transformative, then indispensable. It becomes inefficient, then debilitating, then not credible (even shameful) to continue as before.

The new systems are often opaque, complex, and relentlessly demand modification and updating, in part so they can correctly interact with other evolving technical systems. These attributes heighten users' feelings that they have lost control. As machines grow more capable and more central, resulting anxieties can become existential: Each person worries that his or her worth and individuality are being eroded – worse still, that our species is losing its power and place. Prometheus reached and Icarus flew too far. Dr. Frankenstein created too much.

If only this were all. If only our fears were like those of our 20th-century ancestors in the face of automation or of our 19th-century ancestors in the face of industrialization. But at the end of the second decade of the 21st century – or, if you prefer, the 550th decade of technological progress since the invention of the wheel – something unusual rides alongside these deep-seated,[8] recurring and retrospectively exaggerated fears. It is now not just laymen affected by these technologies who react with unease. It is their creators and proponents. The alarms that concern us are being sounded not by hotheaded Luddites in the factories but rather by a chorus of creators in their laboratories.[9]

It is now not just laymen affected by these technologies who react with unease. It is their creators and proponents.

Expert worry first came to the forefront for American national security policymakers when Einstein, Oppenheimer, and their colleagues created atomic weapons.[10] It is now most pronounced amongst those who work with AI and synthetic biology. Thus Stephen Hawking and his colleagues recorded their view of AI, "Whereas [its] short-term impact . . . depends on who controls it, the long-term impact depends on whether it can be controlled at all."[11] A year later they were succeeded by, at last count, 3,722 "AI/robotics researchers," urging "a ban on offensive autonomous weapons beyond meaningful human control."[12] Biologists worried about breakthroughs in gene editing have raised analogous concerns.[13] The recent Director of National Intelligence responded more sympathetically than his predecessors had to warnings about nuclear weaponry. In Senate testimony he labeled gene editing a potential "weapon of mass destruction":

> Research in genome editing conducted by countries with different regulatory or ethical standards than those of Western countries probably increases the risk of the creation of potentially harmful biological agents or products. Given the broad distribution, low cost, and accelerated pace of development of this dual-use technology, its deliberate or unintentional misuse might lead to far-reaching economic and national security implications. Advances in genome editing in 2015 have compelled groups of high-profile US and European biologists to question unregulated editing of the human germline.[14]

Experts on digital technologies also now speak of catastrophic risk, though for other reasons and in other ways. Here concern arises not so much from uncertainty about what is to come, as from the reality of our present pervasive (and expanding) dependence on a technology so vulnerable to subversion. Experts know how to achieve information integrity, availability, and confidentiality in small systems, but their ability to secure these essentials does not scale. Nonetheless, the ubiquity and proliferation of digital systems, from small "things" to "clouds," is expanding far beyond our control. As one expert put it, "even though security is improving, things are getting worse faster, so we're losing ground even as we improve."[15] Thus another leading thinker began a recent foundational paper by saying, "[W]e state as an axiom that cybersecurity and the future of humanity are now conjoined."[16]

Why Focus on Military Development and Use of Technologies?

Civilian research and development activities often are intertwined with military developments. Even when independently pursued, commercial and academic work – for example, with pathogens – can pose risks of their own. These certainly warrant attention. However, the power and consequence of military weapons demand special attention to technologies related to these systems. Technological developments determine not only the character of these weapons, but also of the sensor, communication, and analytic systems that in large measure determine whether and how weapons are used.

These considerations are joined by another that is less evident, but central to this report. Despite hierarchies of command, rigid procedures, intense training, and elaborate screening of people and plans,[17] military systems have attributes that make them especially prone to human error, emergent effects, misuse, and misunderstanding. These include the secrecy associated with advanced weapons development and use; the unpredictability of operational interactions and environments; the mismatch between experts' skills and military assignments; the interdependencies and vulnerabilities that exist between military systems, especially on the scale of the U.S. military; the urgent deployment of new technologies to meet battlefield operational needs; and finally, the unconstrained nature of military competition.

This section considers each of these in turn.

Secrecy

Civilian uses of complex technologies are commonly subject to review, oversight, and regulation. These are not perfect or cost-free. But when the SEC regulates high speed computer trading, the FDA drug developments, the FAA aircraft innovations, and so forth, they often reduce risk and sometimes puncture bubbles of overconfidence.[18] Visibility also promotes information exchange so that accidents and errors generate insight that supports improvements by developers and operators.[19] Public data permits third parties to evaluate risks, press opposing views through litigation, legislative proposals, publicity that will affect brands, etc. In contrast, military systems have very limited visibility in the United States and much less still in authoritarian nations.

Even inside military organizations, knowledge of classified technologies is frequently kept from those who may be affected by them and even called to use them in wartime or other emergencies. One result is that mainstream technologies are not well coordinated with classified, particularly highly classified, systems. Civilian organizations with complex interactions strive to optimize visibility of their interoperability and interdependencies.[20] Systems of classification, intensified by a widely recognized propensity to overclassification,[21] hide parts of the organization from one another.

The Unpredictability of Operational Interactions and Environments

In a classic analysis 25 years ago, Scott Sagan described "high reliability organizations" as "relatively closed systems in the sense that they go to great efforts to minimize the effects that actors and the environment outside the organization have on the achievement" of safety.[22] Pre-testing, though highly imperfect, is a major way of accounting for, and insulating against, external variables. Accordingly, before we depend on civilian technologies and deploy them at scale, we commonly insist that they be extensively tested in the full range of environments in which they will operate. Self-driving cars are accumulating tens of millions of miles of experience in real-world conditions as a predicate to their normal use. Years of carefully calibrated clinical trials are prerequisites to the introduction of drugs and medical devices. Beta testing in a variety of use environments precedes deployment of major software systems. The military extensively tests its products,[23] but warfare is among the least predictable of endeavors. The timing, locations, and circumstances of use, even the characteristics of adversaries and of users, are highly variable and at best very imperfectly foreseen.[24] The hidden complexities of massive operations[25] are compounded by the incentives for opponents to surprise one another. Accordingly, tests of military technologies and extrapolations to use are probably less reliable than for civilian systems. Accident rates are typically highest in new systems and diminish as experience accumulates and incremental improvements are introduced.[26] Military innovations benefit from the same evolution – for example, with successive generations of fighter aircraft – but military technologies held in reserve, otherwise rarely exercised or hastily introduced,[27] do not have that benefit.

Assignment Policies Often Poorly Match Supervisors and Technologies

Military investments in training are exceptional and some military components – notably, for example, the nuclear Navy – demand technical excellence as a prerequisite to promotion. Generally, however, field grade officers and senior enlisted men and women are rotated

every two or three years among different positions and in and out of different operating and geographic situations. There is much less stability than in civilian environments. The technical environments in which military are assigned and reassigned are large and variable; promotion policies prioritize management breadth and not technical depth; training is frequently not well matched to the challenges of operating both legacy and modern systems in different environments. Accidents result.[28]

Operational Interdependencies

Over the last quarter century, our military has given the highest priority to joint operations and building "systems of systems."[29] As a result, many of our defense systems are exceptionally connected and interdependent even by the standards of an era when networking is widespread in civilian systems.[30] America's strategic commitment, likely to be followed by other militaries, is to move aggressively further in this direction. As a reporter recently observed, "service chiefs are converging on a single strategy for military dominance: connect everything to everything."[31]

These systems are very complex, opaque, interdependent[32] and subject to the "CACE principle: Changing Anything Changes Everything."[33] Operational reliance on the combination of separate systems increases vulnerability to emergent effects. It "creates a strong entanglement: improving an individual component model may actually make the system accuracy worse if the remaining errors are more strongly correlated with the other components."[34] In commercial enterprises, networked systems are heavily policed and partitions strongly disfavored.[35] Networks transcending military services, by contrast, are built across siloed bureaucracies with proud heritages derived from the 18th and 19th centuries.

Urgency and Importance of Uninterrupted Operations

Clichés honor, but civilians barely grasp, the military commitment to mission and a related "can-do" attitude in urgent situations. Peacetime design and acquisition of military weapons and platforms are painstakingly risk averse, but in wartime and even in "peacetime" combat the military commonly accepts consequences that would be intolerable in a civilian setting. During the Second World War, for example, in a rush to make use of an emerging technology, the United States moved from biplanes to jets and, while making this transition, introduced more than 50 types of fighter aircraft and great numbers of variants of bombers and support aircraft.[36] The imperatives of assimilating these new capabilities as quickly as possible overwhelmed concern for safety: During the war, flight accidents in the continental United States "accounted for over 15,000 fatalities, the equivalent of a World War Two infantry division."[37]

The Unconstrained Nature of Military Competition

Commercial competitions between corporations are intense, but military competitions have more at stake: Richly endowed nation-state rivals operate with great paranoia and little inhibition. In military competitions security agencies presume sabotage, indeed proudly practice it. Even botched attempts at sabotage increase the risks of accidents and unintended effects. Moreover, fear of military opponents intensifies willingness to take risks: If they *might* be doing X, we *must* do X to keep them from getting there first, or at least so that we understand and can defend against what they might do.

On March 1, 1954, the United States tested the most potent thermonuclear weapon in its history, resulting in three times the expected yield (in megatons) and unanticipated radioactive fallout. The test, named Castle Bravo, is depicted here. (United States Department of Energy/National Oceanic and Atmospheric Administration)

Causes of Loss of Control

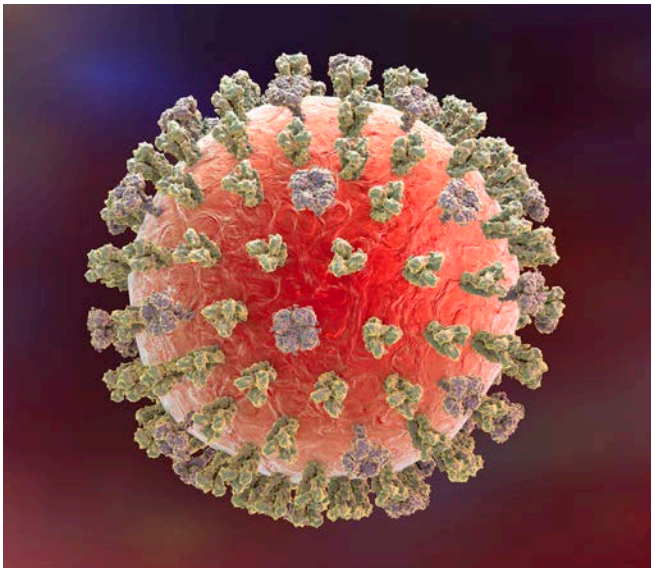
Many are the paths for loss of control of military technology. This section chronicles five primary concerns: analytic errors by technology creators, operational errors by technology users, unintended and unanticipated emergent effects from the evolution and interaction of technologies, sabotage of these technologies by opponents that can lead to malfeasance and errors in decisionmakers' situational awareness and therefore misdirected employment of the technologies upon which they depend.

Analytic and Operational Errors by the United States or Opponents

Military examples of analytic errors include miscalculations before a 1954 atomic bomb test that led to a yield three times greater than anticipated and consequent serious radiation exposure of 665 people.[38] Operational errors by our military include two atomic bombs falling off the wings of a B-52 in flight in 1961,[39] a B-52 crash while carrying four 1.1 megaton bombs in 1968,[40] a failure to remove a training tape indicating nuclear attack so the watch officer thought an attack was occurring and counterattack aircraft were launched before being called back,[41] and the mislabeling and shipment of live anthrax to 86 laboratories in 2015, described by the deputy secretary of defense as "a massive institutional failure with a potentially dangerous biotoxin."[42]

To assess and address the risks of accidents, our national security agencies must consider potential opponents' systems as well as our own. Technology proliferation increases technology risks even if others were as invested in safety as we are. These risks multiply as the number of actors and the number of their interactions grow. Furthermore, these problems are made much worse because others are not likely to be as safe as we are.[43] Many other militaries operate with tighter budgets, more antiquated equipment, fewer controls, and less concern about the human and environmental costs of accidents.[44]

Some accidents will have only local effect as when, for example, in 1979 a Soviet facility inadvertently released anthrax in Svertlovsk, killing some 80 people.[45] Others will have primary consequences locally and secondary consequences globally, as when inadequately trained personnel couldn't cope with an explosion from a badly designed Soviet reactor at Chernobyl in 1986.[46] But some will have as serious a set of consequences for others as they will for the forces that lose control. For example, the prevalent view among experts is that either a Chinese open air vaccine test or a Russian laboratory accident released a pathogen that caused the 1977 worldwide H1N1 pandemic.[47]



Experts believe that the 1977 H1N1 pandemic may have been caused by an open-air vaccine test or laboratory accident that released the pathogen, illustrated here. (Kateryna Kon/Science Photo Library/Getty Images)

Emergent Effects

Emergent effects are attributes that are not identifiable in any individual component of a system but that can be observed in the overall system. Consciousness is an example of an emergent behavior in human beings – we cannot identify it in any component of our bodies, but it "emerges" from those components. As an example of emergence in group behavior, a famous early analysis demonstrated that if all individuals in a neighborhood prefer that 60 percent of their immediate neighbors be racially like them, the neighborhood would move to complete segregation, even though no member of the group sought that outcome.[48] A short story by science fiction writer Arthur C. Clarke provided a dramatic picture of how a fictional military technology ("The Field") could prove disastrous because of emergent consequences:

> The first trial maneuvers proved satisfactory and the equipment seemed quite reliable. Numerous mock attacks were made and the crews became accustomed to the new technique. . . . [However, in combat] there was a hysteretic effect, as it were, and the initial condition was never quite reproducible, owing to all the thousands of electrical changes and movements of mass aboard the ship while the Field was on. These asymmetries and distortions were cumulative, and though they seldom amounted to more than a fraction of one per cent, that was quite enough. It meant that the precision ranging equipment and the tuned circuits in the communication apparatus were thrown completely out of adjustment. . . . Given time, we might even have overcome these difficulties, but the enemy ships were already attacking . . . Our magnificent Fleet, crippled by our own science, fought on as best it could until it was overwhelmed and forced to surrender . . . . This is the true story of our defeat.[49]

In this report, the term "emergent effects" will be used to refer not only to emergence within a single system, but also to effects that emerge when two or more systems interact. Examples of emergent effects in military affairs are not compiled on websites or lists, as they are for accidents.[50] World War I, however, can be seen as an emergent consequence of technological adaptation.[51] In the first decade of the 20th century, European national security planners recognized that they must modernize their mobilization systems to take advantage of the railroad and the telegraph. Each country's resulting "improvements" made sense on its own. But the interactive result of changes in all European nations was to create a system in which a destabilizing event (the assassination of an archduke) that might otherwise have been contained forced multiple European countries into preemptive mobilization. The tail of technology wagged the dog of decisions about war.[52] In the resulting conflagration 20 million people died and 21 million people were wounded.[53]

Though their 2017 analysis does not refer to this precedent, the former third-ranking civilian at the Pentagon and a colleague note the same tendencies from recent and anticipated U.S. and Russian technological developments that increase dependencies on vulnerable space and cyber assets.[54] This situation incentivizes "use it or lose it" first strikes.

Advances in non-nuclear as well as nuclear strategic strike capabilities, and the way they interact, will have a significant impact on the prospects of "slippery slopes" of rapid escalation from crisis to conflict. They will do so singly, but it is particularly their interactions in the context of crisis and early conflict that is of concern. This is especially likely as the actual nature, scope and consequence of the use of such novel technologies may not be clearly anticipated or understood, compounding the already severe "fog of war."[55]

Sabotage and Espionage

The networked attributes of modern technologies make weapons and support systems more accessible than their analog predecessors. The complexity of these technologies renders it easier to hide a sabotage and easier, if discovered, to mask it as a system defect or accident. And close ties between commercial and military systems combine with the global diffusion of science and commercial technology to facilitate reverse engineering.

Among many consequences, three are particularly significant for our purposes. First, military officials who develop new technologies must presume that these technologies will be acquired by others.[56] No warfare technology has been indefinitely exclusively retained. Arguably, first-mover advantages yield sustained

The networked attributes of modern technologies make weapons and support systems more accessible than their analog predecessors.

superiority or at least interim benefits during a period of exclusivity. But first movers need to evaluate more than competitive benefits. They also must weigh these benefits against the risks that arise as their initiatives permeate the global system.[57]

Second, military innovators need to recognize that in any major state conflict, sabotage will be prevalent, so that substantial uncertainty will attend the use of high-technology weapons. Poisoning,[58] misdirecting,[59] and disabling an opponent's systems are part of the stock-in-trade of intelligence and military work. As experience with information system vulnerabilities has indicated, thwarting these efforts is even more difficult than with traditional assets.[60]

In some circumstances, between some opponents, advance recognition of that uncertainty may be stabilizing. It is likely, for example, to inhibit China and the United States from full-scale conflict with one another – neither can be sure of what would happen and who would prevail. Nations with a smaller stake in the status quo or a more desperate sense of vulnerability, however, might find uncertainty acceptable, even desirable.[61] Third, just as sabotage may be misinterpreted as accident, in tense situations accidents may be misinterpreted as sabotage and trigger retaliation.[62]

University of Michigan Ph.D. candidate Kevin Eyholt and a team of researchers demonstrate how the strategic placement of stickers on a STOP sign can distort machine perception. Known as physical adversarial perturbation, this tactic can cause inaccurate or missed object classifications, leading to real-world harm (e.g. an autonomous vehicle algorithm not recognizing a STOP sign). (Kevin Eyholt, et al./University of Michigan)

Misunderstanding by Responsible Officials

Limited understanding hobbles policymakers' discussions and decisions about whether to develop, how to employ, and how much to rely on complex capabilities. Increasingly, senior officials are called to make decisions about, and on the basis of, technologies that did not exist at the time of their education and earlier experience. On-boarding processes do correct these deficiencies. Very few have the time, talent and taste to update their understandings, but most do not. As a result, a former director of the NSA and the CIA writes about his experience in the White House Situation Room debating a use of cyber weapons:

> [T]hese weapons are not well understood by the kind of people who get to sit in on meetings in the West Wing . . . I recall one cyber op, while I was in government, that went awry . . . In the after-action review it was clear that no two seniors at the final approval session had left the Situation Room thinking they had approved the same operation. [63]

We should not be surprised. Sixty years earlier:

> Truman asked Oppenheimer when he thought the Russians would develop their own atomic bomb. Oppenheimer replied . . . he did not know. Truman then said he did know. The answer, he said confidently, was "never." Obviously, Truman had not understood what Oppenheimer had said . . . and what the Los Alamos scientists had tried to tell him.[64]

A colleague recalled, "Truman's statement and the incomprehension it showed just knocked the heart out of [Oppenheimer]."[65]

Inability to Reliably and Persuasively Measure These Risks

It would be desirable to measure or at least reasonably estimate the risks from the just-described problems. Unfortunately, Taleb's observation is sound: "one does not observe the barrel of reality." While highly debatable premises and inferences may provide isolated points of insight, a reasonably informative and reliable assessment of the probability and consequences of these technology risks is beyond current abilities.[66]

Our experience with digital information systems should bring this point home. As America built digital dependencies our national security agencies had only very limited insight into digital vulnerabilities. Experts differed on fundamental questions such as whether vulnerabilities were sparse or dense.[67] They frequently were – and are – surprised by vulnerabilities, even fundamental vulnerabilities, that they did not anticipate.[68] Independent actors have developed stockpiles of vulnerabilities that endure, apparently remaining unknown to defenders who otherwise would patch them.[69] Properly read, the 21st-century experience of digital information technologies is a humbling history of limited understanding and repeated surprise.[70]

The limits of our predictive power are more mathematically and extensively illuminated by the efforts of biologists to assess the likelihood of accidents before establishing research laboratories working with dangerous pathogens. In the United States, Biological Safety Level Three Facilities are required for work with pathogens that "can cause serious or potentially lethal disease through respiratory transmission." Biological Safety Level Four Facilities provide:

> the highest level of biological safety. . . . The microbes in a BSL-4 lab are dangerous and exotic, posing a high risk of aerosol-transmitted infections. Infections caused by these microbes are frequently fatal and without treatment or vaccines. Two examples of microbes worked with in a BSL-4 laboratory include Ebola and Marburg viruses."[71]

The number of these American laboratories grew greatly after the 2001 anthrax attacks. There are now some 1,500 BSL-3 facilities and 15 operating BSL-4 facilities in the United States.[72]

A credible review of a risk assessment for a proposed American BSL-4 facility dealing with agricultural pathogens offered this summary:

> The first Department of Homeland Security risk assessment . . . estimated . . . [an] escape risk [of] over 70% likelihood for the 50-year life of the facility, which works out to be a basic probability of escape . . . = 2.4% per year. The National Research Council overseeing the risk assessment remarked "The . . . estimates indicate that the probability of an infection resulting from a laboratory release of [foot and mouth disease] from [this facility] approaches 70% over 50 years . . . with an economic impact of $9–50 billion. The committee finds that the risks and costs could well be significantly higher than that. . . ." While the DHS subsequently lowered the escape risk to 0.11% for the 50-year lifetime, the NRC committee was highly critical of the new calculations: "The committee finds that the extremely low probabilities of release are based on overly optimistic and unsupported estimates of human error rates, underestimates of infectious material available for release, and inappropriate treatment of dependencies, uncertainties, and sensitivities in calculating release probabilities."[73]

The biologists who offered this summary add their own judgment: "We have more trust in the NRC committee conclusions, as they have no skin in the game." They offer their own calculation that over a ten-year period, ten carefully operated labs will have an 18 percent likelihood of at least one escape from undetected infection of a laboratory worker. If this occurs they observe that there is great variability in whether consequent exposure of others might cause a pandemic – this depends on the rate of reinfection associated with the pathogen – but they assess that the probability of a pandemic in this circumstance is between 1 percent and 30 percent.[74]

A reasonably informative and reliable assessment of the probability and consequences of these technology risks is beyond current abilities.

How would one extrapolate from these assessments to global risks given that a recent survey paints a darker picture of international bio-safety controls, including "gaps in biosafety norms for high-consequence research

that may lead to accidents with pandemic potential, and which should be addressed to increase laboratory safety, worldwide?"[75] If this uncertainty applies to a biotechnology assessment where there is a considerable, well-documented track record, how confident could national security agencies be about risks from the development, testing, or deployment of new systems of AI, autonomous systems, space weaponry, and other technologies, many of them yet to be developed? And if all this could be done, how would they assess the risks from technology theft and proliferation, from potential sabotage, or from interactive effects in peacetime and in war?

In short, U.S. national security leaders should not delude themselves that they have maps for the terrain they are traversing.[76] This intensifies our difficulties in achieving progress – or even consensus about the importance of progress – in reducing risks from new technologies. Efforts for climate control engender substantial debate and resistance, but they operate from a foundation of scientific theory and evidence that helps to distinguish between baseless fear and well-founded fact. They have this advantage because they are dealing with technologies (those associated with burning fossil fuels) that have existed for decades, even centuries, and whose consequent effects can, to some degree, be measured. Even then, they are hotly disputed. The problem is significantly harder when dealing with new technologies whose nature and consequences are necessarily more speculative.

Humans in the Loop

It is commonly thought to be beneficial to require human review of important machine decisions. If a machine action involves a likelihood of loss of life a human role is often said to be essential. This is taken as an axiom by our military establishment. "Defense Department watchers are always keen to remind people that official policy is to keep humans at the top of the command-and-control loop, overseeing – or at least retaining veto power – over the decision to take life."[77]

This position is psychologically empowered: Human beings gain comfort from the concept of human control. Trying to separate our analysis from our feelings, we can identify three considerations that argue for a human role in systems of machine decisions.

The first is that a role for humans adds a different kind of reasoning to a system that relies on machines. If we want to bias a system against error from easy use, an added hurdle is desirable.[78] Accordingly, for example, we require two operators, not just one, to launch a missile with a nuclear warhead. Further security is achieved if the second hurdle is well differentiated from the first. Machine systems can be hacked and misdirected in certain ways, while subversion of humans requires quite different tactics. The two together are less vulnerable to misdirection than one alone.[79]

This perspective, it should be noted, values a human role because it is additive and differentiated, not because it improves the likelihood of decisions later viewed

Google DeepMind's AI program AlphaGo defeated South Korean professional Lee Sedol in 2016. AlphaGo was initially trained with data from human games and then achieved superhuman performance by improving through self-play. A later version, AlphaGo Zero, achieved superhuman performance without any initial human training data. (Kim Min-Hee-Pool/Getty Images)

as correct. A second argument holds that a human adds a transcendent qualitative factor to a machine decision: He or she can take account of things, like context or morality, that programmers have not yet confidently and comprehensively been able to program into machines.[80] Accordingly, a system that empowers human oversight of machines is better protected against undesirable action than a machine system would be alone.

A third argument derives from the unease we feel about unleashing a system that will follow its own programmed priorities without our having the ability to redirect it or shut it down. This concern arises not so much from the intrinsic nature of machines, but from the limits of our abilities to perfectly program them. As the discussion above indicates, there will be errors and unanticipated effects in the design and operation of complex systems. We want a check at the back end to make it less likely that these systems will run amok.

All three of these arguments are valid. However, human decisions are overvalued as a mechanism of control. This is partly because they induce errors of their own, partly because there are many situations where a bias against action is not desirable,[81] and mostly because, far from being an independent or dominant second source of judgment, human decision-makers are machine-dependent to a greater degree than is acknowledged. As a result, though it yields some benefit, the prescription of human intervention is too weak and too frequently counterproductive to consistently control automated systems where speed is at a premium. Moreover, this safeguard can be expected to steadily decline over time. The important lesson for purposes of this report is that efforts to control these systems are vastly more effective at the time of design than they can be in decisionmaking during operations. Early is imperative. Late is too late.

This section will seek to illuminate these points when dealing with AI. Before doing so, however, it is important to note that a number of concerns raised about AI may be relevant as well to other technologies. Biological systems, for example, are likely to become more prevalent in military operations as sensors, means of manufacturing products, and weapons.[82] In these systems, as in AI, opportunities for control, though imperfect, are preponderantly at the front end. Once unleashed, living systems respond to imperatives of their own – they act autonomously in response to environmental stimuli. Whatever means humans have for control arise predominantly from the forethought with which we design biological entities at the outset.

A much-discussed case from 30 years ago provides a useful case study in human/machine decisionmaking in combat. In 1988, officers on board the cruiser USS *Vincennes*, judging that they were under attack, launched two missiles at an Iranian civilian airliner, killing 290 passengers and crew. After extensive review, the U.S. Department of Defense was sympathetic to the crew's error, even decorating the captain and others for their conduct. It did this notwithstanding the central fact that the targeted aircraft was not descending (as an attacker would), but rather ascending as a commercial aircraft would. Grappling with this issue, the official report observed that

> there is a direct contradiction between the data tapes obtained from USS VINCENNES and the situation report submitted by USS VINCENNES . . . following the engagement. . . . Clearly, [the relevant officer] could not have been reporting from the data displayed on the CRO [character read-out]. The most reasonable explanation is contained in the report by [other officers] that his behavior was induced by a combination of physiological fatigue, combat operations, stress and tension which can adversely affect performance and mission execution. As [a prior report] states, "The concept of 'scenario fulfillment' could seem as applying in this case." Since the TIC [Tactical Information Coordinator] has no doubt that the aircraft is an Iranian F-14, heading toward the ship, and is not acknowledging repeated warnings, "the mind may reject incongruent data and facilitate misperception which promote internal consistency." His mental agitation is reflected in his testimony that he took it upon himself to take "every open shot" he was getting on Circuit 15 to ensure "everyone up in the command decision area was informed, kept aware of what was going on in case they got sidetracked on other events." Toward the end it is reported he was yelling out loud.[83]

The *Vincennes* incident "remains a contentious issue"[84] and other factors, including ambiguities in the way the automated system presented data, the ship's experience in other operations, and the participants' personalities, figure into the equation.[85] Our purpose is not to assess responsibility but to assess the role of humans in controlling dangerous technological operations in stressful situations. With five minutes[86] in which to make

Pictured here in 2003, the United States Navy cruiser USS Vincennes *(CG-49), mistaking an Iranian civilian aircraft as an attacking fighter jet, launched two missiles at the aircraft in July of 1988 and killed all 290 aboard. The incident remains an important study in human-machine integration and decisionmaking, due to the USS* Vincennes*' data presentation system and the role of human judgment in the decision to fire the missiles. (Getty Images)*

a life-or-death judgment, the human decisionmakers on the *Vincennes* (a) were buffeted by emotions and distractions, (b) were highly dependent on procedures deducing threat status from machine sensors, and (c) made a fundamental error reporting a critical machine output.

This man-machine relationship has been much considered in commercial aviation, an industry that has seen steady movement toward increased machine capabilities and some competing conclusions about how much room is left for human control, even with highly trained and experienced pilots. "Airbus's philosophy appears to be that automation generally knows better than the pilot and should, under most circumstances, have the final decision authority."[87] Airbus, for example, "defines hard tight envelope limits, beyond which the pilot cannot go regardless of circumstance. By contrast, Boeing sets soft limits that pilots can go beyond if they deem it necessary."[88] Though at present it is a close question which of these two systems is preferable,[89] over time the direction of change seems clear. If, as is very likely, machine decisionmaking improves more substantially than human decisionmaking, the arena of deference to humans will contract.

Processing speed and communication capabilities also increasingly tilt the equation against human decisionmakers, both generally and especially in military situations. Humans now proceed at about one-millionth the speed of machines. Machines are becoming faster. Humans aren't. When instant response is imperative, even our Defense Department's proponents of humans in the loop concede that their desired human control cannot be achieved. It can be anticipated that this exception will swallow the rule as the range of tasks that can be accomplished by machines grows, machine speeds increase (both in calculation and in kinetic operations),[90] and autonomous operations proliferate. As two observers conclude, "military superpowers in the next century will have superior autonomous capabilities, or they will not be superpowers."[91]

Humans now proceed at about one-millionth the speed of machines. Machines are becoming faster. Humans aren't.

A software expert candidly addresses this phenomenon, emphasizing how human control of distant systems also introduces an unacceptable dependency on communications capabilities:

> These are comforting terms such as "semiautonomous" and "human in the loop." However, . . . [e]ffective machine functionality in a variety

> of situations requires full autonomy and a wink and a nod to 'man in the loop' is actually detrimental . . . For example, how do we expect a swarm of autonomous undersea vehicles to act when they have a critical target in sight . . . but realize that communications are being jammed. . . . When a missile enters a two and a half mile radius around a ship, a human doesn't have enough time to react. The Phalanx system must operate in a completely autonomous way. It tracks the missile, aims and fires completely on its own.[92]

With the military and civilian experiences in mind, we should consider the ultimate preserve of human decisionmaking: our president's exclusive authority to order nuclear retaliation. Then-Secretary of Defense Ash Carter emphasized, "Let's not forget that when it comes to using force to protect civilization one of our principles ought to be that there's a human being involved in making critical decisions. I think that's an important principle. . . ."[93] But while Secretary Carter's words trumpet commitment they whisper qualifications. He did not say that "to protect civilization," the president (or even a group of people around the president) makes the critical decisions. He said there was "a human being involved." In fact, a president asked to decide whether to launch upon warning is in a position like that of the captain of the *Vincennes*, with about the same decision time,[94] except that, when compared with a Navy captain, a president has less training and experience relevant to the launch decision he or she might confront. If informed that missiles have been launched against the United States, on what basis would a president decide whether to credit this assessment and respond?

Our president does not look through the windows of the oval office and observe missiles landing. Instead he or she relies on the counsel of a small group of advisors, and they in turn are dependent on a network of sensors, algorithms, and models that conclude (more precisely, we might say, support the inference) that missiles have been launched and are aimed at us. In those minutes our commander-in-chief, like the captain of the Vincennes, relies on the quality of information from the sensors, algorithms, data inputs, and models. About these, our chief executive and those around him or her are likely to have only a small degree of understanding.[95] The most that can be said is that "a human being is involved." What appear to be two separate systems – the machine and the human – are interdependent, and the greater influence on the decision resides in the machinery. This is a common situation in the modern age. Human decisionmakers are riders traveling across obscured terrain with little or no ability to assess the powerful beasts that carry and guide them.[96]

Issues of the human role in machine systems are fair grounds for debate. For our purposes a half dozen conclusions seem reasonable: Human decisionmakers add value because they complicate an attacker's ability to subvert systems; at present, humans can add context that is elusive for machines; the human role will contract as machine capabilities expand, particularly for assessing context[97] and as more decisions need to be made at machine speeds; even at present, decisionmakers are so dependent on machine inputs, they commonly do not add significant independent power as a mechanism of control; to the extent they are empowered, human decisionmakers are as likely to produce unintended consequences as to effect control; urgency stresses and distorts human decisionmaking, but machine complexity confounds human decisionmakers even when there is ample time for considered judgment.

What Is to Be Done?

Analytic and predictive uncertainty and the limits of a human in the loop solution are not a prescription for ignoring a problem. Recognizing our limitations, it is all the more urgent to get a better (though imperfect) understanding of our military technology risks and to establish precautions against them. Each new technology introduced into military systems has unique attributes and risks. However, the technologies addressed here share two particularly troubling characteristics: They are developing faster than our mechanisms of control, and they have the potential for producing especially traumatic, one could say catastrophic, consequences.

Control theory developed to protect us from systems running beyond acceptable parameters.[98] For example, a thermostat measures many familiar systems so that if they run aberrationally hot or cold that out-of-state condition is recognized and steps are initiated to restore equilibrium or shut down the system. Familiar closed loop systems are comfortingly subject to this control. The technologies of concern to us have attributes that are not familiar and have the potential to move beyond the boundaries of closed loops.[99]

They also fall within what Nassim Taleb calls the fourth quadrant.

> In some situations, you can be extremely wrong and be fine, in others you can be slightly wrong and explode. If you are leveraged, errors blow you up; if you are not, you can enjoy life. . . . Some decisions require vastly more caution than others. . . . For instance you do not 'need evidence' that . . . a gun is loaded to avoid playing Russian roulette, or evidence that a thief is on the lookout to lock your door. You need evidence of safety – not evidence of lack of safety.[100]

Our national security agencies have under-invested in comprehending and documenting the problems that may arise in military contexts from new technologies. Recourse to the proposition that they are keeping humans in the loop should offer limited comfort. Humans can add to machine decisionmaking but judgments by military officers are now commonly shaped by machines and machine influence will only increase as machine capabilities grow and demands for speed increase. The required offset is not to put humans "in the loop" or "on the loop" so much as to be maximally thoughtful and creative about new technologies at the time of their design and deployment.

In this light, here are five recommended initiatives, building from the most modest to the most expansive.

1. Increase defense and intelligence agencies' focus on risks of accidents and emergent effects. Do this, for example, by making these concerns a significant part of quadrennial reviews, periodic national intelligence estimates, net assessments, war games, and red team exercises. Nothing is free. This recommended effort will be at the cost of some marginal investment otherwise used to assess and plan for conflict. This report argues, however, that this allocation of resources is warranted: Though the risks of inadvertent actions cannot be well calculated, a better-balanced system would at least educate senior decisionmakers about them and encourage them to take more account of them as urged in the following recommendations. Red team attacks are particularly valuable for exposing vulnerabilities commonly overlooked by program proponents and day-to-day operators.

> **The technologies of concern to us have attributes that are not familiar and have the potential to move beyond the boundaries of closed loops.**

Against this, some might respond that the consequences of intentionally initiated global warfare would be much greater than the consequences, however traumatic, from accident. By this view, accordingly, it would be ill advised to divert any preventive effort from the former to the latter. However, this argument envisions a binary state of badness: on the one hand, accidents that cause tolerable trauma and, on the other hand, malevolent acts that can cause intolerable catastrophe. This distinction is dubious. History and analysis clarify that accidents and emergent effects are significant causes of catastrophic malevolent acts.[101] One value of the recommended games and assessments would be to clarify this and to suggest firebreaks and other ways of mitigating this risk.

There is a second consideration. Agencies' development and deployment of new technologies will be jeopardized if catastrophic accidents occur and, most especially, if these agencies are ill prepared when they do occur. In this light, the recommended initiative, far from undermining technological investment for warfighting and deterrence, is a necessary protection for that investment.

Beyond that, national security agencies need to increase knowledge of the risks these technology investments entail, to declare and dramatize these risks, and to stimulate processes that will be called upon in the event of accident.[102]

Some top-level policy guidance is a prerequisite to balancing our interests, and this is most likely to result from heightened attention in regular policy reviews. In addition, at least some documents that reflect this recognition of risk must be public and engage Congress, the White House, and other relevant executive branch agencies and offices. The natural tendency within the national security establishment is to minimize the visibility of these issues and to avoid engagement with potentially disruptive outside actors. But this leaves technology initiatives with such a narrow a base of support that they are vulnerable to overreaction when accidents or revelations occur. The intelligence agencies should have learned this lesson when they had only weak public support in the face of backlash when their cyber documents and tools were hacked.

The natural tendency within the national security establishment is to minimize the visibility of these issues and to avoid engagement with potentially disruptive outside actors.

Leadership attention and broader support alone are not sufficient. Leaders also must foster a more extensive understanding of these risks. The 2008 financial crisis catalyzed work to improve our understanding of how one bank's actions can trigger reactions throughout the whole banking system.[103] The 2013 Ebola crisis stimulated analogous work on epidemic risk.[104] The previous section described our shortfalls in understanding our risks from new technologies and observed that it is hard to assess these risks precisely because the technologies are, and will be, new. War games and simulations will build our understanding.

2. Routinely consider and aim to reduce risks of proliferation, adversarial behavior, accidents, and emergent behaviors when developing, deploying, and employing technologies for warfare. Give priority to reducing these risks. During operations, the role of human decisionmakers will be limited. As technology advances, this role will contract. The most important opportunities for control are before they are used in conflict, that is, when programs and safeguards are designed and implemented. As a celebrated historian observed in other contexts, "the decisive choice is seldom the latest choice in the series. More often than not, it will turn out to be some choice made relatively far back in the past."[105]

The precept is broadly understood. Before new weapons systems are deployed, well-established defense department systems of test and evaluation require a "designated approval authority" to assess the risks of accidents from these weapons. A proponent of cybersecurity innovations, for example, must identify opportunities for adversarial attack and consider these risks, as well as risks of accident, before a new system is introduced.[106] But this approach needs to be applied to emergent risks that do not lend themselves to convincing probabilistic calculation. It should be made applicable to the whole suite of new technologies (biology, robotics, etc., not just IT), to programs in development as well as near deployment, and to assessment of likely costs to us when adversaries adopt technologies we develop and exploit vulnerabilities we present as we increase the range and complexity of our systems.

Encouraged in part by this report, DARPA and IARPA have begun to explore how they might systematically build these considerations into their programs. An appendix to this report presents questions developed by IARPA Director Jason Matheny as prerequisites to be answered before the introduction of new programs.

While American military and intelligence agencies progress with all technologies on which our national security depends, including particularly destructive technologies, it is important that their leaders understand the limits of our comprehension of what they are introducing and that agencies design for resilience in the face of failure.[107] This requires learning from simulations of performance under a variety of conditions, including attack; abstracting from analogous systems;[108] developing data by controlled experience with new systems;[109] auditing those systems[110] as they are maintained,[111] updated, and operated in practice; and assisting and training human operators and supervisors so that they can intervene and if necessary terminate machine operations if warranted.

U.S. national security agencies have models of this approach from the controls with which we have surrounded nuclear technology as a means of propulsion for our submarine force, as weaponry for our strategic nuclear forces, and as a core of energy from our nuclear power plants. Our future uses of new technologies need to be at least as strong as these past programs in their training regimens, procedures, and safeguards. But many of the new technologies are more opaque, more rapidly evolving, and less tightly controlled by human operators than in the past.[112] Furthermore, past programs like those for nuclear submarines and nuclear weapons have been relatively isolated and therefore easier to protect. Risks of sabotage and emergent effects from interaction will be greater with systems that have more autonomy and operate in more populated and interactive technological and physical environments. Consequently, these systems need to be even better understood and better controlled than their predecessors. A United Nations group concerned with lethal autonomous weapons has posed four relevant questions:

- How are existing systems verified (was it built right) and validated (was the right system built)? Are existing and planned autonomous systems scrutable (what do you know and how do you know it)? Can machines describe their learning?
- How are human and social safety issues such as hacking and privacy being tackled? Can autonomous machines be made foolproof against hacking?
- Can there be software/hardware locks on machine behavior, and could a learning machine be prevented from bypassing/changing them?
- Does the transformative character of AI and its possible ubiquity limit the [lethal autonomous weapon systems] discussion in any manner, or is AI like other dual-use technologies in the past?[113]

Two DARPA programs are suggestive about technical possibilities. A first effort recognizes that opaque algorithms must be understood by operators "if future warfighters are to . . . appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners."[114] The "Explainable AI (XAI) program aims to create . . . [n]ew machine-learning systems [that] will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future."[115] A second program pursues control in biological contexts. The Safe Gene Program aims to develop genetic turn-off switches,[116] counter-measures and other tools that "build in biosafety for new biotechnologies at their inception."[117]

3. Independently and regularly assess the risks of accidents and emergent effects from our military technologies. Encourage the same effort from our opponents and credible third parties. The previous section indicated why it is extremely difficult for anyone, no matter how technologically astute, to discern and assess the risks associated with new technologies. It is exponentially harder for those within organizations devoted to developing these technologies to sustain potentially disruptive analyses and to force responses that may minimize risk but reduce effectiveness or raise costs.[118] Two of the most eminent thinkers about "normal accidents"[119] emphasized this variable. Scott Sagan decried the military propensity in accidents he investigated to "circle the wagons."[120] By contrast, Charles Perrow attributed progress in commercial aviation safety in large measure to the "rich mix of interested formal organizations checking upon each other."[121]

In this light, it would be valuable, perhaps invaluable, to supplement organizational assessments with reviews by experts outside proponent organizations.[122] To test this proposition, an external review should be conducted of a sample of perhaps a half dozen programs developing nascent technologies. If this review proves useful, then it might be regularized under the aegis of one of the existing external bodies, for example the Office of Technology Policy in the White House, the National Institute of Standards and Technology, the President's Intelligence Advisory Board, the Defense Science Board, or an Inspector General.[123]

Experience with such reviews should be assimilated before judgments are made about how frequently and robustly they should be undertaken. If of limited value, they might be at substantial intervals or not at all. If of notable value they might be conducted at regular intervals or more or less continuously.

If American agencies found this to be desirable, other nations should be encouraged to establish similar systems. A third party (for example, the United Nations)[124] also might be encouraged to make such assessments and, optimally, to coordinate with national efforts so as to sensitize all nations about risks that can be understood at least in unclassified contexts. National risk analyses and risk reduction efforts could be encouraged by sharing some civilian technology applications with those who complied.[125]

4. Increase multilateral planning with our allies and our opponents so that we are cooperatively better able to recognize and respond to accidents, catastrophic terrorist events, and unintended state conflicts. American policymakers have long understood that it does us little good to safeguard our nuclear weapons if our opponents do not safeguard theirs. They need to extrapolate from that lesson. For example, safeguards surrounding American experimentation with genetic modifications avail us little if modifications created in other countries escape into the global environment. The point is generalizable across all potent technologies. It would be dangerously naive for us to assume that foreign parties carefully assess and evaluate risks in their frame of reference just as we do in ours.

All nations employing advanced destructive technologies need to share perceptions of risks and safeguards.[126] The previous recommendation would help in that regard. From that foundation, we need to plan together to cope with unanticipated and undesired events.[127] A terrorist attack with a contagious pathogen in Moscow or an accident from a laboratory in China will put our security interests very much in play. Conversely, they have a stake in our activities. We need to plan for robust cooperation across as well as within alliance relationships.[128]

To achieve this, national security agencies from many nations should collaboratively use the tools of analysis, simulation, and gaming that have just been recommended to deepen American understanding. These efforts would provide a foundation for cooperation. Risks cannot be understood simply from our perspective, nor can we safely leave others to understand them only from their perspectives.[129]

5. Use the new technologies as means for encouraging and verifying treaties and norms. Deterrence is the big stick that keeps nations from crossing some critical lines. However, American national security officials have supplemented deterrence with a web of treaties and norms that create disincentives to development, deployment, and employment of some technologies.[130] Most evidently, both our success and our failure with these tools are reflected in the fact that there are approximately 15,000 nuclear weapons in the arsenals of nine nations.[131] Though the number is distressingly high, nonproliferation treaties have contributed to keeping it lower than it otherwise would be.[132] And though strategists celebrate deterrence as the primary means of preventing conflict between nuclear states, a norm of no-first use[133] undoubtedly has contributed to keeping nuclear nations from using these weapons against states without nuclear weapons.[134]

This model of treaties and norms has been employed with imperfect but nonetheless great benefit in efforts to control other new military technologies. With varying degrees of success nations have moderated the development, proliferation, and use of missiles,[135] blinding lasers,[136] land mines,[137] weapons in outer space,[138] chemical weapons,[139] etc.[140] Recognizing this, biologists,[141] leaders in AI,[142] cyber experts,[143] and those concerned with other new technologies[144] are attempting to encourage emergent norms[145] and treaty restraints on development, proliferation, and destructive use of these technologies.

Unfortunately, the strength and attraction of these tools is largely proportional to participating nations' faith in our abilities to observe their violation.[146] Nations have made substantial headway with norms and treaties when, and only when, they can find means of monitoring tests, materials, equipment, weapons, and use.[147] Responsible leaders cannot be Pollyanna-ish about possible adversaries who may publicly foreswear technology developments but secretly continue with them.[148] President Reagan's proposition "trust, but verify," seems still to obtain with its strong implicit corollary: When we can't verify, we don't trust.

Our central problem is that by all present measures the technologies described in this report are of low visibility. Biological work that once required industrial facilities and controllable supplies now can be conducted inside of commonplace buildings with commercially available micro-equipment and material.[149] Work on AI is even more difficult to trace as computational hardware has steadily miniaturized and proliferated, and the most basic medium of progress, software, is extraordinarily difficult to access and scrutinize.

There may be opportunities inherent in the very technologies of concern, however. AI systems are particularly valuable as means for detecting activities otherwise lost in the noise of the benign everyday activity. Well-designed biological systems can be exquisite sensors of perturbations in their environments. Information technologies create opportunities for looking within opponents' systems. None of these mechanisms are perfect, but all are powerful. As these technologies develop, American national security officials should make it a priority to broaden and intensify our investments in them as means of arms control, not just armament.

Conclusion

There is good reason for America to pursue technological superiority in the future as it has in the past. As our national security leaders strain in that direction, they must also recognize that even a persistently first-place position in this race entails risks. This essay highlights the point that superiority is not synonymous with security: There are substantial risks from the race.

This would be more tolerable if the race had a foreseeably good ending or if the risks were readily ascertainable, small and controllable. But, unless the paranoia of the human animal abates, the race appears to be unending. The source of that problem is not in our technologies, it is in us. As E. E. Morison observed a half century ago, "[o]ne of the things you can learn from history is that men have lived with machinery at least as well as, and probably a good deal better than, they have yet learned to live with one another."[150]

But technology makes our situation worse. Many have seen how the power of technology amplifies the power of malevolent acts and therefore the risks that nations, groups, and individuals consciously initiate one another's destruction. This report highlights the probability that we don't have to will this result – it may emerge because nations, groups and individuals can't well control what they create. To return to the metaphor with which this report began, we are loading more and more bullets into guns that are more widely distributed not only within our system and among our allies, but also among opponents who are following our technological lead. The resulting technical-human system always has been dangerous, but with each year it is accumulating more risks and the likelihood is that in the future it will be riskier still. America is not alone responsible for this technological roulette, but because it is superior it has the primary responsibility and the primary opportunity to diminish risk. Doing so is imperative to protecting our security.

At the outset, this report observed that many fears about new technologies root more in emotion than in reason. This makes it too easy to discard even well-grounded concerns as irrational or Luddite. Hopefully, this report rebuts that easy dismissal. While reinforcing the case for care, we should not neglect, however, the relevance of emotionally rooted reactions. At a minimum we should recognize that catastrophic accidents are likely to be significant because of the psychological and political reactions they induce as well as for their more immediate material consequences. (Thus, for example, some countries cut back all nuclear programs after radiation was released when a tsunami hit reactors in Fukushima, Japan.)

More positively, fear can be our friend. To be sure, it can paralyze, mislead and feed irrational desires to reverse the arrow of time. But if properly considered, it may alert all users – us and our opponents – to faults in what we are doing and where we are going. Technological developments are empowering, infatuating, and frequently irresistible. But they carry real risks that are obscured by complexity, bureaucratic momentum, misunderstanding, and the thirst that individuals and organizations have for enhanced capabilities. These risks are not effectively controlled by, as some would have it, "keeping humans in the loop." By highlighting the limits of our control, this report tries to strip away false comfort. Fears sensitize us to the risks we commonly overlook and unwisely discount.

Properly acknowledged fear can be a powerful bonding agent. It joins love and self-interest as an essential part of the trinity that motivates people. It is even more important in international relations because nations, to put it mildly, are short on love for one another.

National security officials often identify allies and adversaries as though these were immutable categories carrying necessary consequences in our relationships. But history teaches a different lesson, well illustrated by America on the eve of World War II. Our affinities with the democracies of Western Europe and our yet deeper special relationship with Britain did not motivate us to intervene when Hitler attacked them in 1940 and 1941. But when the United States came to fear for itself, it cemented an alliance even with Stalin's Soviet Union. No opponent we now face is so evil and repugnant as Stalin. We cooperated with him – and him with us – because we saw a greater risk from Hitler and the Nazi state.

It is a well-established trope in science fiction to envision an alien threat as a catalyst that welds together nations that were previously hostile. We don't have to wait for the aliens. If humanity comes to recognize that we now confront a great common threat from what we are creating, we can similarly open opportunities for coming together. Cooperation in this area may become a beachhead for broader cooperation. Put another way, a clear-eyed view of militaries as inadvertent actors may open a path to taming our capacities for destruction. Perhaps this is too much to hope for. But then, perhaps it is not.

Appendix

Questions Developed For Program Managers Making New Proposals by Dr. Jason Matheny, Director, IARPA

1. What is your estimate about how long it would take a major nation competitor to weaponize this technology after they learn about it? What is your estimate for a non-state terrorist group with resources like those of al Qaeda in the first decade of this century?
2. If the technology is leaked, stolen, or copied, would we regret having developed it? What if any first mover advantage is likely to endure after a competitor follows?
3. How could the program be misinterpreted by foreign intelligence? Do you have any suggestions for reducing that risk?
4. Can we develop defensive capabilities before/alongside offensive ones?
5. Can the technology be made less prone to theft, replication, and mass production? What design features could create barriers to entry?
6. What red-team activities could help answer these questions? Whose red-team opinion would you particularly respect?

Endnotes

1. Military technological advantage derives from research initiatives, assimilation of research undertaken by others (including civilians and military competitors), translation of inventions into instruments relevant to warfare, development of strategies and operational concepts that take advantage of these instruments, and providing incentives and training for service members to employ these innovations. A brief informative layman's description of the different aspects of this process is provided by Malcolm Gladwell, "Creation Myth: Xerox PARC, Apple, and the Truth About Innovation," *The New Yorker*, May 16, 2011, https://www.newyorker.com/magazine/2011/05/16/creation-myth. This report refers to all these strands without attempting to disentangle their respective roles.

2. It is also relevant that China's population is three times the size of the United States. This is both an advantage (for example, by demanding technology sharing as a predicate to access to its markets) and a disadvantage (for example, by creating imperatives for greater per capita social expenditures rather than military investments).

3. And supporting systems of command and control, sensors, analytic algorithms, etc.

4. The Future of Humanity Institute at Oxford University has focused attention on these issues, beginning with a 2008 conference at which participants rated the likelihood of a "technologically induced global catastrophe" during the next century at greater than 10 percent. Nick Beckstead et al., "Unprecedented Technological Risks," (Future of Humanity Institute, 2014), https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf, provides a good summary, referencing several technologies, of risks of concern to FHI.

5. Two classics are Scott D. Sagan, *The Limits of Safety: Organizations, Accidents and Nuclear Weapons* (1993) and Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (first published in 1984, reissued with additional material in 1999; all references in this report are to the later edition). Eric Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident and The Illusion of Safety* (2013) provides a modern update focused on a particular accident. Analysis of high reliability organizations has continued in a number of contexts, notably including through the High Reliability Organizations Project led by Todd LaPorte at the University of California at Berkeley. Some of the history of the Berkeley effort and of opposing viewpoints on these issues is chronicled in Mathilde Bourrier, "The Legacy of the Theory of High Reliability Organizations: An Ethnographic Endeavor," (Sociograph – Working Paper n°6 /2011), https://www.unige.ch/sciences-societe/socio/files/4814/0533/5881/sociograph_working_paper_6.pdf.

6. Subversive efforts within the competition will increase these risks.

7. Calestous Juma, *Innovation and Its Enemies: Why People Resist New Technologies* (2016) 294, presents a series of engaging case studies. He emphasizes that "[w]hat appears on the surface as conservatism or irrational rejection of new ideas may represent a deeper logic of social stability woven around moral values, sources of legitimacy and economic interests."

8. Perhaps these fears are inherently human, and our technology anxieties are presaged by our fears of natural disasters and wrath of the Gods. By this view, as technology has secured greater control of the natural environment, perhaps it has taken on the burden of worry previously associated with natural risks, as though there were some psychological law of conservation of fear. This argues that we should be on guard for the play of our emotions. It does not, however, demonstrate that there are no risks.

9. And those laboratories, as Michael Hopmeier has pointed out, are today's factories.

10. "The physics community has a special relationship to nuclear weapons policy. Physicists invented and refined nuclear weapons and historically have made major contributions to efforts to limit the dangers they pose." Steve Fetter et al., "Nuclear Weapons," *Physics Today*, April 2018, 39.

11. Stephen Hawking et al., "Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough?'" *Independent*, May 1, 2014.

12. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," (July 28, 2015) https://futureoflife.org/open-letter-autonomous-weapons/. Follow-up discussion and an articulation of research priorities in 2017 can be found at "Asilomar AI Principles," https://futureoflife.org/ai-principles/.

13. See, for example, Marc Lipsitch and Thomas V. Inglesby, "Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens," (American Society for Microbiology, 2014), doi: 10.1128/mBio.02366-1412 December 2014 mBio vol. 5no. 6 e02366-14: "[R]esearch that aims to create new potential pandemic pathogens (PPP) . . . represents a tiny portion of the experimental work done in infectious disease research, [but] it poses extraordinary potential risks to the public." Leading geneticist Kevin Esvelt says, "I occupy a strange position . . . I am probably the foremost critic of genetic engineering and yet I am also someone at the forefront of the work I critique." Kristen V. Brown, "This Scientist is Trying to Stop a Lab-created Global Disaster," *Splinter*, June 27, 2016, https://splinternews.com/this-scientist-is-trying-to-stop-a-lab-created-global-d-1793857858, quoting Kevin Esvelt, head of MIT's Sculpting Evolution Laboratory.

14. James R. Clapper, "Statement for the Record Worldwide Threat Assessment of the US Intelligence Community (Senate Armed Services Committee, February 9, 2016), 9 (emphasis added), https://www.dni.gov/files/documents/SASC_Unclassified_2016_ATA_SFR_FINAL.pdf.

15. Julie Bort quoting Bruce Schneier, "Now We Must 'Pledge Allegiance' To Apple Or Google To Stay Safe," *Business Insider*, November 7, 2012, https://www.schneier.com/news/archives/2012/11/now_we_must_pledge_a.html.

16. Daniel E. Geer Jr., "A Rubicon," (Hoover Institution, Aegis Series Paper No. 1801, February 2018). Geer continued: "There are two—precisely and only two—classes of cyber risks that directly rise to the level of national security . . . One class is those critical services which by the very definition of their mission must create a single point of failure. . . . The other risk at the level of national security is that of cascade failure which implicates services that are critical or merely pervasive."

17. The military is concerned, of course, to control its members as well as its instruments of warfare. The human aspect of this equation is as important as technology issues, but is not the subject of this report. For readers who wish to consider the imperative of human controls, a compelling starting point might be Jim Frederick, *Black Heart: A Platoon's Descent into Madness in Iraq's Triangle of Death* (2012).

18. See, for example, "In the Matter of Knight Capital Americas LLC: Order Instituting Administrative and Cease-And-Desist Proceedings," October 16, 2013, 2–3, in which the SEC describes its response to automated trading: "Recent events and Commission enforcement actions have demonstrated that this investment must be supported by an equally strong commitment to prioritize technology governance with a view toward preventing, wherever possible, software malfunctions, system errors and failures, outages or other contingencies and, when such issues arise, ensuring a prompt, effective, and risk-mitigating response. The failure by, or unwillingness of, a firm to do so can have potentially catastrophic consequences for the firm, its customers, their counterparties, investors and the marketplace. The Commission adopted Exchange Act Rule 15c3-52 in November 2010 to require that brokers or dealers, as gatekeepers to the financial markets, 'appropriately control the risks associated with market access, so as not to jeopardize their own financial condition, that of other market participants, the integrity of trading on the securities markets, and the stability of the financial system.'"

19. Perrow, *Normal Accidents*, 371–72, 382, distinguishes between error inducing and error avoiding organizations. He identifies information sharing as an important characteristic that played an important role in determining whether a given system was of one or the other type. In this context he observes that the civilian aircraft industry was helped greatly by an extensive record system and "the rich mix of interested formal organizations checking upon each other."

20. See for example, the efforts of a developer infrastructure group charged with increasing visibility of interdependencies in different parts of Google's operations. J. D. Morgenthaler et al., "Searching for Build Debt: Experiences Managing Technical Debt at Google" (Proceedings of the Third International Workshop on Managing Technical Debt, 2012), https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37755.pdf.

21. On the subject generally, see Federation of American Scientists, "What is Over-Classification?" October 21, 2013, https://fas.org/blogs/secrecy/2013/10/overclass/.

22. Sagan, *The Limits of Safety,* 17.

23. Praising "the cautious, conservative Pentagon processes widely derided as obstacles to innovation [the Deputy Secretary of Defense emphasized the importance of] 'operational test and evaluation, where you convince yourself that the machines *will* do exactly what you expect them to, reliably and repeatedly.'" Sydney J. Freedberg Jr., "War Without Fear: DepSecDef Work On How AI Changes Conflict," *Defense One*, May 31, 2017, https://breakingdefense.com/2017/05/killer-robots-arent-the-problem-its-unpredictable-ai/.

24. Clausewitz's famous formulation still applies. War, he observed, "involves tremendous friction which cannot, as in mechanics, be reduced to a few points, is everywhere in contact with chance, and brings with it effects that cannot be measured, just because they are largely due to chance." Karl von Clausewitz, *On War*, edited by Michael Howard and translated by Peter Paret (1989), 120.

25. Steven Johnson, *Emergence: The Connected Lives of Ants, Brains, Cities* (2001), 19, offers the insightful comment: "Complexity characterises the behaviour of a system or model whose components interact in multiple ways and follow local rules, meaning there is no reasonable higher instruction to define the various possible interactions. . . . A complex system is . . . characterised by its inter-dependencies, whereas a complicated system is characterised by its layers."

26. "The lowest sustained fatal accident rate of first generation [Airbus] jets was around 3.0 accidents per million fights, whilst for the second generation it was around 0.7, meaning a reduction of fatal accidents of almost 80% between generations. In comparison, third generation jets now achieve about 0.2 accidents per million fights, a reduction of around a further 70%. Finally, fourth generation jets have the lowest accident rate of all, at a stable average rate of about 0.1 fatal accidents per million fights, which is a further 50% reduction compared to the third generation." See, Airbus, "A Statistical Analysis of Commercial Aviation Accidents 1958-2016." The International Air Transport Association's "Safety Report 2016" (April 2017) proclaims that "[o]ver the last ten years the world's commercial aviation system industry has improved its overall safety performance by 54%," p. 6, https://www.skybrary.aero/bookshelf/books/3875.pdf.

27. Contrast the civilian record described in the previous footnote with the military record discussed in Part II of this report.

28. As a recent example: "In many cases, deferred modernization has delayed the installation of modern integrated Bridge systems. Sailors from one ship cannot simply cross to another ship of the same class and expect familiar equipment or lay-outs. Low cost simulators do not sufficiently provide the fidelity to fully replicate the stresses of complex operations in high traffic areas." Department of the Navy, "Comprehensive Review of Recent Surface Force Incidents," October 26, 2017, 14.

29. William A. Owens, "The Emerging U.S. System-of-Systems" (February 1996), was an early influential statement by the Vice Chairman of the U.S. Joint Chiefs of Staff.

30. A decade ago an astute observer saw this development as "inevitable" in both commercial and military systems and added: "As the expectations for and potential benefits of systems of systems grow, so does the demand for such systems. Their number will continue to increase and their importance to intensify. Nowhere is this acceleration more obvious than in the U.S. Department of Defense, where there is a rising advocacy for transformation, driven by technological advances in computing and communication and instantiated in a vision of system of systems known as network-centric warfare (NCW)." See David A. Fisher, "An Emergent Perspective on Interoperation in Systems of Systems," Carnegie Mellon University Technical Report, March 2006, 9, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.955.5917&rep=rep1&type=pdf.

31. Patrick Tucker, "The Future the US Military is Constructing: a Giant, Armed Nervous System," *Defense One*, September 26, 2017,http://www.defenseone.com/technology/2017/09/future-us-military-constructing-giant-armed-nervous-system/141303/. Tucker supports his observation with quotations from every service chief. For example, "as Adm. John Richardson, Chief of Naval Operations, put it . . . in July, "I want to network everything to everything." For an early promulgation of this approach see Admiral William A. Owens, "The Emerging U.S. System-of-Systems," (National Defense University Strategic Forum, February 1996).

32. Geer observes that "complexity hides interdependence(s), ergo complexity is the enemy of security." "A Rubicon" Hoover Institution, Aegis Series Paper No. 1801 (2018), 2. Geer refers to Steven Johnson, *Emergence: The Connected Lives of Ants, Brains, Cities, and Software* (2001), 19, to explain that "complexity characterises the behaviour of a system or model whose components interact in multiple ways and follow local rules, meaning there is no reasonable higher instruction to define the various possible interactions."

33. David Sculley et al: "Hidden Technical Debt in Machine Learning Systems," https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

34. Ibid.

35. For a description of some efforts in one organization, see J. David Morgenthaler et al., "Searching for Build Debt: Experiences Managing Technical Debt at Google" (2012), https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37755.pdf.

36. A list may be found at Stephen Sherman, "World War Two Aircraft Specs of Fighter Planes by Model and Type" (May 2002, updated January 2012), http://acepilots.com/planes/specs.html.

37. Marilyn R. Pierce, "Earning their Wings: Accidents and Fatalities in the United States Army Air Forces During Flight Training in World War II," Unpublished PhD Thesis, Kansas State University, 2013, Abstract 208. The "can do" attitude is hardly just relevant to the history recounted by Pierce. The U.S. Navy devoted a section to this topic in its 2017 review of accidents in the Pacific Fleet. Among other causes of an unacceptably high accident rate, the review recorded the fact that "the rigor in conducting risk-to-force assessments was overshadowed by operational primacy." Department of the Navy, "Comprehensive Review of Recent Surface Force Incidents," October 26, 2017, 102.

38. Thomas Kunkle and Byron Ristvet, "CASTLE BRAVO: Fifty Years of Legend and Lore," Defense Threat Reduction Agency (D-TRIAC SR-12-001, January 2013), 127 https://web.archive.org/web/20140310004623/http://blog.nuclearsecrecy.com/wp-content/uploads/2013/06/SR-12-001-CASTLE-BRAVO.pdf. The 23-man crew of a Japanese fishing boat that "by some odd coincidence . . . was not detected during pre-shot aerial searches" was also exposed.

39. Broken Arrows: Nuclear Weapons Accidents," http://www.atomicarchive.com/Almanac/Brokenarrows_static.shtml, recites this along with 31 other incidents: "January 24, 1961: While on airborne alert, a B-52 suffered structural failure of its right wing, resulting in the release of two nuclear weapons. One weapon landed safely with little damage. The second fell free and broke apart near the town of Goldsboro, North Carolina. Some of the uranium from that weapon could not be recovered. No radiological contamination was detectable in the area." Collisions between foreign and U.S. aircraft, ships, and submarines are well documented and reflected on this site. See also Patricia Lewis et al., "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy" (2014), 1: "Since 1945 there have been disturbing near misses in which nuclear weapons were nearly used inadvertently. Evidence from many declassified documents, testimonies and interviews suggests that the world has

been lucky." http://www.europeanleadershipnetwork.org/medialibrary/2016/02/04/d2106a19/Is%20Trident%20safe%20from%20cyber%20attack.pdf. See also Schlosser, *Command and Control*.

40. Sagan, *The Limits of Safety,* 156ff., 238 ff.

41. The circumstances of this 1979 event are described by former Secretary of Defense William J. Perry in his memoir *My Journey at the Nuclear Brink* (2015) 52–3.

42. "Deputy Secretary of Defense Robert Work said in a media briefing Thursday[:] 'We are shocked by these failures. DOD takes full responsibility for the failures.'" See, Heath Druzin, "DOD: Live anthrax sent to labs was 'a massive institutional failure,'" *Stars and Stripes*, July 23, 2015, http://www.stripes.com/news/us/dod-live-anthrax-sent-to-labs-was-a-massive-institutional-failure-1.359510. Extensive investigate reporting by *USA Today* on this subject is collected at http://www.usatoday.com/topic/9ee9e5de-b702-4fbc-9e5d-1b595adcf938/biolabs/. More generally, see the federal government's review of 291 regulated biological laboratories that reported "12 potential losses and 233 potential releases" – none assessed to have caused harm to humans – in 2015. See, U.S. Department of Health and Human Services et al., "2015 Annual Report of the Federal Select Agent Program" (June 2016), http://www.selectagents.gov/resources/FSAP_Annual_Report_2015.pdf;GAO, "HIGH- CONTAINMENT LABORATORIES: Improved Oversight of Dangerous Pathogens Needed to Mitigate Risk," https://www.gao.gov/assets/680/679392.pdf; and "CDC 90 Day Internal Review of the Division of Select Agents and Toxins," http://www.cdc.gov/phpr/dsat/documents/full-report.pdf. "The Internal Review Workgroup acknowledged uncertainties and gaps in understanding how best to strengthen biosafety and to implement measures that appropriately balance the ability to conduct life-saving research with biological select agents and toxins and the need to ensure the safety and security of the public and the workers in these institutions."

43. In his memoir *State Secrets: An Insider's Chronicle of the Russian Chemical Weapons Program* (2009), 196–7, 202, Vil. Z. Mirzayanov describes for example the "Devil-may care attitude towards safety of the head of the chemical weapons degasification department . . . the ignorance or carelessness of its employees . . . [and] outdated safety procedures." In general, "The Soviet military-chemical complex was also running their factories without proper waste water and exhaust air treatment." Global variations in civilian aviation accidents (arising for a variety of reasons in a variety of contexts) should alert us to likely differences between American and foreign safety records. Accidents occur several times more often per commercial mile flown in Africa, for example, than in the United States. See, for example, International Civilian Aviation Organization, "State of Global Aviation Safety" (2013, referring to 2012 data), 16, https://www.icao.int/safety/State%20of%20Global%20Aviation%20Safety/ICAO_SGAS_book_EN_SEPT2013_final_web.pdf. The 2014 report, using 2013 data, clusters countries in "regional aviation safety groups" and shows an accident rate for the sub-Saharan African group as four and a half times that for the Americas. International Civilian Aviation Organization, "State of Global Aviation Safety" (2014, referring to 2013 data), 8, Appendix II defining the regions: https://www.icao.int/safety/Documents/ICAO_2014 Safety Report_final_02042014_web.pdf; see the 2016 report at https://www.skybrary.aero/bookshelf/books/3681.pdf.

44. It should be noted that the United States doesn't always have the most constraining sensitivities. For example, "judging by opinion polls and government policies, most Western democracies, with the possible exception of France, have been more anti-nuclear than the United States." See, Nina Tannenwald, "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use," *International Organization*, 53 no. 3 (Summer 1999), 464.

45. Jeanne Guilleman, *Anthrax: The Investigation of a Deadly Outbreak* (1999). An open-air release of a smallpox virus that caused ten deaths is described in Milton Leitenberg and Raymond A. Zilinskas, *The Soviet Biological Weapons Program: A History* (2012), 129–131.

46. "One person was killed immediately and a second died in hospital soon after as a result of injuries received. Another person is reported to have died at the time from a coronary thrombosis. Acute radiation syndrome (ARS) was originally diagnosed in 237 people onsite and involved with the clean-up and it was later confirmed in 134 cases. Of these, 28 people died as a result of ARS within a few weeks of the accident. Nineteen more subsequently died between 1987 and 2004 but their deaths cannot necessarily be attributed to radiation exposure. Nobody offsite suffered from acute radiation effects although a large proportion of the childhood thyroid cancers diagnosed since the accident are likely to be due to intake of radioactive iodine fallout. Furthermore, large areas of Belarus, Ukraine, Russia, and beyond were contaminated in varying degrees." World Nuclear Association, "Chernobyl Accident, 1986" (updated November 2016), http://www.world-nuclear.org/information-library/safety-and-security/safety-of-plants/chernobyl-accident.aspx.

47. The DNA of the 1977 virus was nearly identical to that in a 1950 outbreak – an event extremely unlikely to occur in nature where viruses mutate frequently in response to the pressure of natural selection. Debate over the cause of this unlikely situation could be assembled as a case study in the difficulties of biological attribution. The predominant view has been that "the virus was frozen in a laboratory freezer since 1950, and was released, either by intent or accident, in 1977. This possibility has been denied by Chinese and Russian scientists, but remains to this day the only scientifically plausible explanation." "Origin of Current Influenza H1N1 Virus," Virology Blog, March 2, 2009, http://www.virology.ws/2009/03/02/origin-of-current-influenza-h1n1-virus/. See also R. G. Webster et al., "Evolution and Ecology of Influenza A Viruses," *Microbiology Review*, 56 nos. 152–179 (1992), 170; and Marc

Lipsitch and Alison P. Galvani, "Ethical Alternatives to Experiments with Novel Potential Pandemic Pathogens," *PLOS Medicine*, May 2014, http://journals.plos.org/plosmedicine/article/asset?id=10.1371/journal.pmed.1001646.PDF/. The most recently published review of the evidence concluded: ". . . the tripartite origin of the outbreak in northeast China that produced almost identical isolates is not supportive of the conclusion that this was a single laboratory accident. It is more likely that either the vaccines produced from these stocks or the viruses themselves used in tests of vaccine development were virulent enough to spark the 1977 epidemic. The bulk of the evidence rests with this possibility: the unnatural origin, mildness of presentation of the virus, widespread dissemination of cases in a short amount of time, temperature sensitivity of the samples, contemporary observations, and existence of live-virus vaccine trials which were occurring at that time." See Michelle Rozo and Gigi Kwik Gronvall, "The Reemergent 1977 H1N1 Strain and the Gain-of-Function Debate" (July/August, 2015) mBio 6(4):e01013-15. doi:10.1128/mBio.01013-15. A credible third view argues that the data is so sparse and subject to error that neither the time nor the source of initial cases can be identified with confidence. See Joel O. Wertheim, "The Re-Emergence of H1N1 Influenza Virus in 1977: A Cautionary Tale for Estimating Divergence Times Using Biologically Unrealistic Sampling Dates," *PLOS One*, June 2010, e11184. Prof. Terry Leighton of Oakland Children's Hospital, who identified several of these references, comments, "It is clear that fine grained phylogenetic and phylo-geographic datasets [which are not available] are required to constrain competing hypotheses for outbreak sources." Email to the author, April 2, 2018. On problems of attributing biological attacks, see Anne L. Clunan et al., *Terrorism, War or Disease* (2008).

48. The phenomenon, demonstrated originally by Tom Schelling, is well explained by Frank McCown, "Schelling's Model of Segregation," http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/. Joshua Epstein and Robert Axtel, *Growing Artificial Societies* (1996), used a simulation of interactions of cells containing sugar to rigorously demonstrate this phenomenon. It is accordingly often referred to as a sugarscape agent model.

49. Arthur C. Clarke, "Superiority" (1951), http://www.mayofamily.com/RLM/txt_Clarke_Superiority.html.

50. Interactions between intelligent machines in commercial settings could give us other examples. Mark Wilson, "AI Is Inventing Languages Humans Can't Understand. Should We Stop It?" (July 14, 2017), https://www.fastcodesign.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it, discusses languages invented by Facebook computers to more efficiently communicate with one another. Google's David Sculley et al. observe: "[H]idden loops may exist between completely disjoint systems. Consider the case of two stock-market prediction models from two different investment companies. Improvements (or, more scarily, bugs) in one may influence the bidding and buying behavior of the other." "Hidden Technical Debt in Machine Learning Systems," https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

51. This is not to say that technology was the only or even primary cause. The point is that emergent effects contributed to the outbreak of a conflict that was not sought by decision-makers amongst the combatant nations.

52. The classic account is Barbara Tuchman, *The Guns of August* (1962). A more recent history, Christopher Clark, *The Sleepwalkers: How Europe Went to War in 1914* (2012) offers insight on railroads on 305, 308, 322, 336, 352, and 420.

53. Nadège Mougel, "World War I Casualties" (REPERES), http://www.centre-robert-schuman.org/userfiles/files/REPERES%20%E2%80%93%20module%201-1-1%20-%20explanatory%20notes%20%E2%80%93%20World%20War%20I%20casualties%20%E2%80%93%20EN.pdf.

54. James N. Miller, Jr. and Richard Fontaine, "A New Era in US-Russian Strategic Stability: How Changing Geopolitics and Emerging Technologies are Reshaping Pathways to Crisis and Conflict," (Harvard University, Belfer Center, 2017).

55. Ibid., 16.

56. "RMA [Revolution in Military Affairs] proponents and skeptics alike assume that innovations will spread. They differ in their assessment of the ease and speed with which this will occur." Leslie C. Eliason and Emily O. Goldman, "Theoretical and Comparative Perspectives on Innovation and Diffusion," in Leslie C. Eliason and Emily O. Goldman (eds.), *The Diffusion of Military Technology and Ideas* (2003), 6.

57. Of course, it will be relevant to consider whether others may go down a development path as fast (or possibly even faster) even if we restrain our investment. The Russian offensive biological weapons program appears to have intensified when we announced that we would not pursue such weapons, in part because our announcement was taken as a ruse. See generally, Milton Leitenberg and Raymond A. Zilinskas, *The Soviet Biological Weapons Program: A History* (2012).

58. Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (Future of Humanity institute, 2018) observes at pp. 17–18: Today's AI systems suffer from a number of novel unresolved vulnerabilities. These include data poisoning attacks (introducing training data that causes a learning system to make mistakes), adversarial examples (inputs designed to be misclassified by machine learning systems), and the exploitation of flaws in the design of autonomous systems' goals. These vulnerabilities are distinct from traditional software vulnerabilities (e.g. buffer

overflows) and demonstrate that while AI systems can exceed human performance in many ways, they can also fail in ways that a human never would."

59. "'The concern about this is that one might find that an adversary is able to control, in a big-data environment, enough of that data that they can feed you in misdirection,' said Dr Deborah Frincke, head of the Research Directorate (RD) of the US National Security Agency/ Central Security Service (NSA/CSS). Adversarial machine learning, as Frincke called it, is 'a thing that we're starting to see emerge, a bit, in the wild. It's a path that we might reasonably believe will continue,' she said." See Stilgherrian, "Machine learning can also aid the cyber enemy: NSA research head,"(Full Tilt, March 15, 2017), http://www.zdnet.com/article/machine-learning-can-also-aid-the-cyber-enemy-nsa-research-head/. Ian Goodfellow et al. provide a good account of poisoning machine learning in "Attacking Machine Learning with Adversarial Examples" (February 24, 2017), https://blog.openai.com/adversarial-example-research/. See also Nicolas Papernot et al., "Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples," arXiv preprint arXiv:1602.02697 (2016): "attackers [can] control a remotely hosted [deep neural network] with no access to the model, its parameters, or its training data [assuming] that the adversary can observe outputs from the target [deep neural network] given inputs chosen by the adversary. We introduce the attack strategy of fitting a substitute model to the input-output pairs in this manner, then crafting adversarial examples based on this auxiliary model."

60. Stealing the technology for the atomic bomb was challenging. Stealing it for the Air Force's most advanced fighters or appropriating considerable numbers of the NSA's cyber attack tools was apparently less so. Department of Justice Press Release, "Chinese National Pleads Guilty to Conspiring to Hack into U.S. Defense Contractors' Systems to Steal Sensitive Military Information," March 23, 2016, https://www.justice.gov/opa/pr/chinese-national-pleads-guilty-conspiring-hack-us-defense-contractors-systems-steal-sensitive.

61. I discuss this point in Richard Danzig, "Surviving on a Diet of Poisoned Fruit: Reducing The National Security Risks of America's Cyber Dependencies" (Center for a New American Security, 2014).

62. After noting that both Russia and the United States "possess strong incentives to use cyber and counter-space capabilities early in a conflict to gain advantage," Miller and Fontaine ("A New Era") continue: "This emerging situation could greatly increase the risks of stumbling into conflict due to accident or inadvertence." Sketching a scenario for major cyber conflict around a South China Sea incident, Jonathan Reiber and Arun Mohan Sukumar envision: "The PLA covertly penetrates the networks of Cho Ray Hospital in Ho Chi Minh City for purposes of surveillance and collection on key leaders. During the surveillance operation, a young PLA major inadvertently tests malware that destroys data regarding the timing and delivery of medicine to two wards of patients. In the first known deaths caused by a data-disruptive cyberattack, four patients die at Cho Ray hospital from failing to receive adequate amounts of medication through their medical devices."Reiber and Sukumar anticipate escalatory effects when American forensic analysis determines the source of these attacks, but the U.S. intelligence community cannot ascribe the motive for them. "Asian Cybersecurity Futures: Opportunity and Risk in a Rising Digital World" (Center for Long Term Cyber Security, December, 2017), 37.

63. Michael V. Hayden, *Playing to the Edge: American Intelligence in the Age of Terror* (2016), 147. In correspondence about this report, Paul Scharre suggested that decision-makers' unfamiliarity may bias them toward seeing the benefits, but not possible adverse consequences, of new technologies.

64. Ray Monk, *Robert Oppenheimer: A Life Inside the Center* (2012), 494.

65. Monk, *Robert Oppenheimer,* 494.

66. Lee Clarke, *Mission Improbable: Using Fantasy Documents to Tame Disaster* (2003) is, as its title indicates, fiercely assertive about this point. See also the supportive discussion of his work in Perrow, 374ff. Issues about the limits of calculation are important in many fields. Peter Lewin, "What Do We Know for Certain About Uncertainty," (Keynote Remarks Presented at The Legatum Institute Charles Street Symposium, June 2012) provides a good discussion about the centrality of this issue in economics.

67. See Bruce Schneier, "Should U.S. Hackers Fix Cybersecurity Holes or Exploit Them?" *The Atlantic*, May 19, 2014, https://www.theatlantic.com/technology/archive/2014/05/should-hackers-fix-cybersecurity-holes-or-exploit-them/371197/.

68. For the most recent example, see Peter Bright, "As predicted, more branch prediction processor attacks are discovered," *Ars TechnicaU*, March 26, 2017: "[These attacks] demonstrate that the isolation boundaries that have long been assumed to exist are rendered somewhat permeable by the speculative execution hardware that is essential to high-performance processors. Moreover, BranchScope shows that Spectre isn't the only avenue through which this speculative execution can be exploited. . . . It's likely to be years before researchers have determined all the various ways in which the speculative execution hardware can be used to leak information this way, and it will be longer still before robust, universal defenses are available to stop the attacks."

69. "[A]ny serious attacker can always get an affordable zero-day for almost any target . . . zero-day vulnerabilities have an average life expectancy of 6.9 years." See Lily Ablon and Andy Bogart, "Zero Days, Thousands of Nights: The Life and Times of Zero Day Vulnerabilities and Their Exploits," (RAND, 2017), xiii.

70. "[I]n some important respects, the leaders of American government respond to the problems of cyber technology as two hundred years ago Americans and their leaders responded to our Western frontier. We are excited by what we already have experienced and believe that future discoveries and exploitations will have immense and transformative effects. At the same time, we are uncertain and conflicted – what is the shape of the new territory? How should it be governed? Must it be insecure for decades to come? How do we reconcile old values, interests, power relationships, and practices with the frontier's unfamiliar risks, demands, and ways of doing things?" Richard Danzig, "Forward," to Richard M. Harrison and Trey Herr, *Cyber Insecurity: Navigating the Perils of the Next Information Age* (2016), xi.

71. Centers for Disease Control and Prevention, "Quick Learn Lesson: Recognizing the Biosafety Levels," https://www.cdc.gov/training/quicklearns/biosafety/.

72. The number of BSL 4 facilities is cited in Keith Rhodes (Chief Technologist, GAO Center for Technology and Engineering), "High-Containment Biosafety Laboratories: Preliminary Observations on the Oversight of the Proliferation of BSL-3 and BSL-4 Laboratories in the United States," (October 4, 2007). We don't seem to be able to reliably count the number of BSL-3 laboratories in the United States. The best calculation was that there were 1,495 registered BSL-3 laboratories in 2010. Jocelyn Kaiser, "Taking Stock of the Biodefense Boom," *Science*, September 2, 2011, 1214. It is notable that even the GAO appears to rely on this reporter's estimate drawn from registrations eight years ago. GAO commented that "there is no reliable source of the total number of high-containment laboratories in the United States." General Accountability Office, "High-Containment Laboratories: Assessment of the Nation's Need Is Missing," February 25, 2013, 6, https://www.gao.gov/assets/660/652308.pdf.

73. Lynn C. Klotz and Edward J. Sylvester, "The Consequences of a Lab Escape of a Potential Pandemic Pathogen," *Front Public Health*, no. 2 (2014), 116, 10.3389/fpubh.2014.00116, PMCID: PMC4128296.

74. Ibid.

75. Gigi Kwik Gronvall and Michelle Rozo, "Synopsis of Biological Safety and Security Arrangements," http://www.upmchealthsecurity.org/our-work/pubs_archive/pubs-pdfs/2015/Synopsis%20of%20Biological%20Safety%20and%20Security%20Arrangements%20UPMC%20072115.pdf.

76. Continuing the analogy to our exploration of the frontier, "Like our forbearers we can only partially comprehend and map the unknown [terrain of cyber security]. We do so relying on reports filtering back from scattered settlements and from explorers who tell us, sometimes inaccurately, where some trails lead. With haphazard information, we try to regulate areas that are at least somewhat settled, attempt occasional forays to maintain a modicum of order in some critical areas, and more or less acquiesce to anarchy in the remainder. Barely able to comprehend what will come, we project our hopes and fears onto our mental maps of the future." Richard Danzig, *Cyber Insecurity*, xi.

77. Patrick Tucker, "The Future the US Military is Constructing: a Giant, Armed Nervous System," *Defense One*, September 26, 2017, http://www.defenseone.com/technology/2017/09/future-us-military-constructing-giant-armed-nervous-system/141303/. Twenty years ago, an academic observer captured the same thought: "The most durable pocket of military tradition is the belief that human beings are ultimately responsible and ultimately in control, until and unless their equipment fails them," in Gene I. Rochlin, *Trapped in the Net: The Unanticipated Consequences of Computerization* (1997), 167.

78. Of course, we also can program this bias into machines. When IBM demonstrated the power of its AI system, Watson, by having it vanquish human competitors at Jeopardy, "Watson could be tuned to be either more aggressive in answering questions (and hence more likely to be wrong) or more conservative and accurate." Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age* (2014), 26.

79. For these reasons, I have argued that a strong presumption should be created that critical systems integrate non-cyber safeguards in their access and operation. Non-cyber components, also described as "out- of-band" measures, can include, for example, placing humans in decision loops, employing analog devices as a check on digital equipment, and providing for non-cyber alternatives if cyber systems are subverted. Digital information systems that drive information technologies import only digital vulnerabilities. We can protect ourselves by forcing attackers to cope with system attributes that are outside the reach of computer code. Richard Danzig, "Surviving on a Diet of Poisoned Fruit: Reducing the National Security Risks of America's Cyber Dependencies,"(Center for a New American Security, July 2014), 21. Dan Geer has argued more generally for retaining analog systems as complements to critical digital systems. Daniel E. Geer, Jr., "A Rubicon" (Hoover Institution, Aegis Series Paper No. 1801, February 2018). Peter Levin additionally observes, in correspondence with the author, that while one common vulnerability may undermine an elaborate system of, for example, digital machines, a single tactic is less assured of success in dealing with human decisionmakers because each will have his or her own idiosyncrasies.

80. This contextual understanding seems to have been central to the American watch officer's discounting of the training tape incident described, and the celebrated decision of a Soviet officer who declined to report an incoming nuclear attack when machine output appeared to suggest it was underway.

81. In combat, for example, we fear false negatives (mistaking an attacker for an innocent actor) as much as false positives (mistaking an innocent actor for an attacker).

82. Many nations, including the United States, have foresworn biological weapons. After making such a declaration, the Soviet Union continued with a vast biological weapons program. The Soviet program is documented in Milton Leitenberg and Raymond A. Zilinskas, *The Soviet Biological Weapons Program: A History* (2012).

83. U.S. Navy, Formal Investigation into the Circumstances Surrounding the Downing of Iran Airflight 655 on 3 July 1988, Investigation Report, 13.

84. Nancy C. Roberts, and Kristen Ann Dotterway, "The *Vincennes* Incident: Another Player on the Stage?" *Defense Analysis*, 11 no. 1 (1995), 31.

85. "[T]here are many unanswered questions about the *Vincennes* incident as the various hearings, reports and articles attest." Roberts and Dotterway, 42. Roberts and Dotterway observe that investigators may have overstated the human failure: "Finding a discrepancy between system data and crew's recollections, and unable to explain it, the investigators speculated that the problem had to be due to task fixation, scenario fulfillment and combat-induced stress. Perhaps, given the pressure of time, it was easier to find fault with the operators than to leave a critical or embarrassing question unanswered or to challenge the efficacy of tactical support systems such as NTDS (Link 11)."Roberts and Dotterway, 43. See also David Evans, "*Vincennes*: A Case Study," U.S. Naval Institute *Proceedings*, August 1993, https://www.usni.org/magazines/proceedings/1993-08/vincennes-case-study. Scott Sagan also wrote about this instance as a case study in the challenges of establishing rules of engagement. Sagan emphasized that the Navy's rules of engagement were liberalized after a previous incident, involving the USS *Stark*, had resulted in that ship being victimized by an attack. Scott D. Sagan, "Rules of Engagement," *Security Studies*, 1991, 78–108, https://doi.org/10.1080/09636419109347458.

86. U.S. Navy, Airflight 655 Investigation Report, 4.

87. Clint R. Balog, "Airbus v. Boeing: Whose Automation Philosophy is Best?" *Aerospace America*, July-August 2015, 33ff.

88. Ibid., 34.

89. Balog prefers Boeing's position. "While computers may be gaining on the human brain, it remains the most capable, flexible and adaptable complex processor in the world." Balog, "Airbus v. Boeing," 35. William Langewiesche also focuses on the role of the human pilot in his contrasting accounts of two Airbus accidents in 2009. The first, *Fly by Wire* (2009), celebrated a pilot's achievement in managing a safe US Airways landing in New York's Hudson River after his engines became disabled by a bird strike. The second describes compounding pilot errors leading to the deaths of 228 people on an Air France flight from Rio de Janeiro to Paris. "The Human Factor," *Vanity Fair*, October 2014, https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash. Pages 98–115 of *Fly by Wire* provide an account of the evolution of "fly-by-wire" digital technologies, including the observation that "the new technology was revolutionary. . . . For the first time in history, airplanes could be made that would be fundamentally different from the Wright brothers Flyer." Balog, "Airbus v. Boeing," 9.

90. For example, cruise variants of hypersonic missiles (missiles flying at greater than five times the speed of sound) likely to enter inventories over the next decade could "hold targets within a 1,000km radius . . . at risk and could strike these targets within several minutes." Generally, "defenders with capable terrestrial and space sensors will have only a few minutes to know these missiles are inbound." Richard H. Speier et al., "Hypersonic Missile Nonproliferation: Hindering the Speed of a New Class of Weapons" (RAND, 2017), 12, 14.

91. Cara LaPointe and Peter L. Levin, "Automated War: How to Think About Intelligent Autonomous Systems in the Military," *Foreign Affairs*, September 5, 2017. It should be noted that the authors hedge their assertion with the phrase "in the next century." As with many such predictions, the trend is clear but the timing is uncertain. In this author's view it will not take more than one or two decades to validate this statement.

92. Amir Husain, *The Sentient Machine: The Coming Age of Artificial Intelligence* (2017) 100–101.

93. Nicholas Thompson, "The Former Secretary of Defense Outlines the Future of Warfare," *Wired*, February 19, 2017, https://www.wired.com/2017/02/former-secretary-defense-outlines-future-warfare/?mbid=social_gplus. See also David Emery, "Robots with Guns: The Rise of Autonomous Weapons Systems," Snopes, April 21, 2107, https://www.snopes.com/2017/04/21/robots-with-guns: "Current Deputy Secretary of Defense Robert Work (who also held that position under President Obama) is a firm believer in AI on the battlefield, and also takes a strong stand against excluding humans from the loop. He . . . dismissed talk of 'killer robots' and comparisons to *The Terminator*. 'Humans, in the United States' conception, will always be the ones who make decisions on lethal force, period,' he said. 'End of story.'"

94. Jeffrey Lewis, "Is Launch Under Attack Feasible?" Nuclear Threat Initiative, August 24, 2017, http://nti.org/6687A, offers an invaluable walk through the decisionmaking process, including his judgment that "The U.S. President is left with at best 2–3 minutes to weigh options and consider alternatives." Decision-times of single-digit minutes such as faced the commander of the *Vincennes* and would face the president are becoming commonplace as the speed of weapons increases. Richard H. Speier, "Hypersonic Missile Nonproliferation," 12, 14.

95. This has been widely observed even amongst operators with more technical training and everyday experience. Describing Nadine Sarter as "an industrial engineer at the University of Michigan, and one of the pre-eminent

researchers in the field," William Langewiesche reports: "Sarter has written extensively about 'automation surprises,' often related to control modes that the pilot does not fully understand or that the airplane may have switched into autonomously, perhaps with an annunciation but without the pilot's awareness. Such surprises certainly added to the confusion aboard Air France 447. . . . Sarter said, 'We now have this systemic problem with complexity, and it does not involve just one manufacturer. I could easily list 10 or more incidents from either manufacturer where the problem was related to automation and confusion. Complexity means you have a large number of subcomponents and they interact in sometimes unexpected ways. Pilots don't know, because they haven't experienced the fringe conditions that are built into the system. I was once in a room with five engineers who had been involved in building a particular airplane, and I started asking, 'Well, how does this or that work?' And they could not agree on the answers. So I was thinking, if these five engineers cannot agree, the poor pilot, if he ever encounters that particular situation . . . well, good luck.'" See William Langewiesche "The Human Factor," *Vanity Fair*, October 2014, https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash.

96. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2016) provides an illuminating survey of the power of algorithms and how often their premises are uncritically accepted in everyday life. Christian Sandvig et al., "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association, May 22, 2014, provides a good survey of algorithm biases likely to be invisible to users.

97. The question of how humans add benefit could be avoided when humans were clearly better than machines at logical thought and analysis of contextual clues. However, this foundation has been shaken by the steady accumulation of machine victories over humans in games such as chess, go, Jeopardy, and poker. These are specific tasks, but ". . . even before the recent burst of progress, this portion has expanded steadily over time. In addition, it has often been the case that once AI systems reach human-level performance at a given task (such as chess) they then go on to exceed the performance of even the most talented humans. Nearly all AI researchers in one survey expect that AI systems will eventually reach and then exceed human-level performance at all tasks surveyed. Most believe this transition is more likely than not to occur within the next fifty years." See Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation" (Future of Humanity Institute, 2018), 16.

98. For an informative history see James R. Beniger, *The Control Revolution: Technological and Economic Origins of the Information Society* (1986).

99. The Safe Gene program is an effort to bring control theory to genetic engineering. For a general discussion, see Domitilla Del Vecchio et al., "Control Theory Meets Synthetic Biology," (Royal Society Publishing, June 2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971224/pdf/rsif20160380.pdf.

100. Nassim Nicholas Taleb, "The Fourth Quadrant: A Map of the Limits of Statistics," https://www.edge.org/conversation/nassim_nicholas_taleb-the-fourth-quadrant-a-map-of-the-limits-of-statistics.

101. See the discussion of World War I and the comments of Miller and Fontaine about the risks of accidentally induced space and cyber conflict quoted in Part of II of this report.

102. When an unprecedented tsunami in 2011 created a near meltdown of the second of Japan's two Fukushima nuclear plants, the United States played a central role in support of Japan's response. But while we could jury-rig responses that drew on extensive planning for U.S.-Japanese actions in the event of war, we had almost no contingency planning for an accident and had to improvise command and communication relationships, transport flows, etc. See Richard Danzig and Andrew Saidel, "Beyond Fukushima: A Joint Agenda for U.S. and Japan Disaster Management,"(CNAS, 2012 (https://www.cnas.org/publications/reports/beyond-fukushima-a-joint-agenda-for-u-s-japanese-disaster-management. As another example, in 2013 and 2014, U.S. leaders recognized strong American interests in countering Ebola in West Africa. U.S. national security agencies played a central role in Liberia and supported Britain and France in their efforts in Guinea and Sierra Leone. But the consequences of an outbreak would have been as grave, and U.S. security agencies requirements far more demanding, if the outbreak occurred in states that did not welcome Western intervention.

103. Stefano Battiston et al. provide a brief and good summary in "Complexity theory and financial regulation: Economic policy needs interdisciplinary network analysis and behavioral modeling, *Science*, February 19, 2016, http://polymer.bu.edu/hes/rp-battiston16.pdf.

104. Hans Heesterbeek et al., "Modeling Infectious Disease Dynamics in the Complex Landscape of Global Health," *Science*, March 13, 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445966/.

105. Arnold J. Toynbee quoted at AZQuotes.com, http://www.azquotes.com/quote/945879.

106. Department of Defense Instruction 8500.01, Subject Cyber-Security, March 14, 2014, https://rmf.org/images/stories/rmf_documents/850001_2014.pdf; Department of Defense Instruction 8510.01, Subject: Risk Management Framework (RMF) for DoD Information Technology (IT), March 12, 2014, https://rmf.org/images/stories/rmf_documents/851001_2014.pdf.

107. This is not to say that all users must be experts. Rather, it emphasizes that we need to identify and continuously engage some experts with deep understanding of important systems we are using and that those who utilize these systems should have a working understanding of the limits and faults of the technologies that they are employing.

108. For example, IARPA has initiated "The Functional Genomic and Computational Assessment of Threats (Fun GCAT) program . . . to develop new approaches and tools for the screening of nucleic acid sequences, and for the functional annotation and characterization of genes of concern, with the goal of preventing the accidental or intentional creation of a biological threat." It aims to develop "tools [that] will enhance the ability to computationally and functionally analyze nucleic acid sequences [and] ascribe threat potential to known and unknown genes through comparisons to the functions of known threats." IARPA-BAA-16-08, September 2016, https://www.iarpa.gov/index.php/research-programs/fun-gcat.

109. "How do we expect autonomous systems, especially those capable of supporting or executing lethal action, to decide on their own? How do we imbue them with judgment consistent with our values? Basically, the same way we train personnel: by leveraging our knowledge about past events and previous situations in order to predict future outcomes." Cara LaPointe and Peter Levin, "Automated War: How to Think About Intelligent Autonomous Systems in the Military," *Foreign Affairs*, September 5, 2016, https://www.foreignaffairs.com/articles/2016-09-05/automated-war. See also O'Neil, *Weapons of Math Destruction*, 209: We can improve algorithms "only when we have an ecosystem with positive feedback loops."

110. Observing that "[m]athematical models should be our tools, not our masters," O'Neil *Weapons of Math Destruction*, 207–8, calls for "algorithmic audits." Sam Arbesman, "Overcomplicated Technology And The Need For Biological Thinking: Complex Systems Like Stock Trading Software Need To Be Studied Like An Ecosystem," *Ars Technica*, November 6, 2016, http://arstechnica.com/business/2016/11/overcomplicated-technology-and-the-need-for-biological-thinking/ suggests: "Field biologists are supremely aware of the assumptions they are making, and know they are looking at only a sliver of the complexity around them at any one moment. . . . Similarly, when encountering a complicated tangle of a technological system . . . a physics mind-set will take us only so far if we try to impose our sense of elegance or simplicity upon its entirety. If we want to understand our technological systems and predict their behavior, we need to become field biologists for technology."

111. A classic account of the complexities of software systems observes that "[f]ixing a defect has a substantial (20–50 percent) chance of introducing another. So the whole process is two steps forward and one step back. . . . All repairs tend to destroy the structure, to increase the entropy and disorder of the system. . . . 'Things are always at their best in the beginning,' said Pascal." Frederick OP. Brooks, *The Mythical Man-month: Essays on Software Engineering* (1975, Anniversary Edition 1995), 122–3.

112. Consider for example the risks of attack on additive manufacturing. "[N]o commercially available technology visually ensures that the printed layers match what the designer intended. Although some nascent research in implementing machine vision to monitor a build has emerged, it currently only checks the printed part against the STL file. If attackers modify the STL file or the NC code for printing the part, the printer will report that the layer is correct even if it differs from the designer's intentions because it's checking against the modified file." Hamilton Turner et al., "Bad Parts: Are Our Manufacturing Systems at Risk of Silent Cyberattacks?" www.computer.org/security.

113. "Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Food-for-thought paper Submitted by the Chairperson," September 4, 2017, https://www.unog.ch/80256EDD006B8954/(httpAssets)/2117A10B-536751D2C1258192004FD7EA/$file/FoodforthoughtPaper_GGELAWS_Final.pdf.

114. David Gunning, "Explainable Artificial Intelligence (XAI)," (2016) http://www.darpa.mil/program/explainable-artificial-intelligence. The Director of IARPA, Jason Matheny, urged developers of AI systems to "please bake-in explainability from the start" because doing that was a predicate to making the outcomes of these systems persuasive to policymakers. https://www.youtube.com/watch?v=z-vpdFWPYBs (presentation of Jason Matheny at 23:20.) He argued that "explainability" was worth its cost even if it diminished effiency because it enhanced usability.

115. Gunning, "Explainable Artificial Intelligence (XAI)." Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, April 11, 2017, comments more generally: "Starting in the summer of 2018, the European Union may require that companies be able to give users an explanation for decisions that automated systems reach. This might be impossible, even for systems that seem relatively simple on the surface, such as the apps and websites that use deep learning to serve ads or recommend songs. The computers that run those services have programmed themselves, and they have done it in ways we cannot understand. Even the engineers who build these apps cannot fully explain their behavior." A recent path-breaking paper provides incidental insight into one way that explainable AI can work. After demonstrating that machines could predict homosexuality from complete facial images, investigators developed insight into the basis of machine decisions by feeding it partial information (e.g., quadrants of faces). Systematic variation revealed key variables that the machine relied on for differentiation. Yilun Wang and Michal Kosinski, "Deep Neural Networks Can Detect Sexual Orientation from Faces," *Personality*

and Social Psychology, 2017, https://osf.io/fk3xr/. See also Andrew Lohn et al., "How We Can Overcome the Risks of AI," *Time*, October 22, 2015, http://time.com/4080577/artificial-intelligence-risks/.

116. The Wyss Institute describes its efforts along these lines in Benjamin Boettner, "Kill Switches for Engineered Microbes Gone Rogue," November 16, 2017, https://wyss.harvard.edu/kill-switches-for-engineered-microbes-gone-rogue/. Stuxnet has such an off-switch, terminating the program at a prescribed date. Note, however, that there are always uncertainties. In the case of Stuxnet, termination was dependent on a correct setting of time and date in a using computer. Andrew Leedom, "Stuxnet: Risk & Uncertainty in the First Salvo of Global Cyber Warfare," *The SAIS Europe Journal,* April 1, 2016, http://www.saisjournal.org/posts/stuxnet.

117. "'The field of gene editing has been advancing at an astounding pace, opening the door to previously impossible genetic solutions but without much emphasis on how to mitigate potential downsides," said Renee Wegrzyn, the Safe Genes program manager." DARPA's press release continues: "DARPA launched Safe Genes to begin to refine those capabilities by emphasizing safety first for the full range of potential applications, enabling responsible science to proceed by providing tools to prevent and mitigate misuse. Each of . . . seven teams will pursue one or more of three technical objectives: develop genetic constructs – biomolecular "instructions" – that provide spatial, temporal, and reversible control of genome editors in living systems; devise new drug-based countermeasures that provide prophylactic and treatment options to limit genome editing in organisms and protect genome integrity in populations of organisms; and create a capability to eliminate unwanted engineered genes from systems and restore them to genetic baseline states." See DARPA, "Building the Safe Genes Toolkit," https://www.darpa.mil/news-events/2017-07-19. See also Joseph Garthwaite, "U.S. Military Preps For Gene Drives Run Amok: DARPA Researchers Are Developing Responses For Accidental Or Malicious 'Genetic Spills,'" *Seattle Times*, November 18, 2016.

118. Lynn Eden, *Whole World on Fire: Organizations, Knowledge and Nuclear Weapons Devastation* (2004) provides an illuminating history of how the U.S. Air Force devoted considerable attention to blast effects from nuclear weapons, but largely ignored fire consequences likely to extend two to five times farther than blast damage. She draws the general lesson that organizations strive for "solutions to problems that those in the organization have decided to solve. . . . The potential for mishap is clear . . . physical processes may be poorly understood or unanticipated," (pages 3, 5). Organizational routines and choices were self-reinforcing: "the effects of learning, high research costs, interdependence and self-reinforcing expectations reinforce choices already made. For blast damage, expert knowledge was encoded into routines that continually built more organizational capacity to predict blast damage. For fire damage, expert knowledge was not translated into organizational routines, and predictive capacity was not built. . . . Perhaps most significant, participants made sense of the increasing divergence of organizational capability to predict blast damage and incapability to predict fire damage by understanding it to demonstrate that fire damage could not, in fact, be predicted," (pages 285–286).

119. See endnote 5.

120. Sagan, *The Limits of Safety*, 237.

121. Perrow, *Normal Accidents*, 382.

122. Discussing development of potentially pandemic pathogens, two experts write: "Given the stakes in this process, the risk assessment process should be directed by those without a clear personal stake in the outcome, just as peer review of science is performed by those without a direct interest in the outcome. The credibility of the risk assessment will depend both on the rigor of the quantitative process and the perceived objectivity of the process." Marc Lipsitch and Thomas V. Inglesby, "Moratorium on Research Intended to Create Novel Potential Pandemic Pathogens," (American Society for Microbiology, 2014), doi: 10.1128/mBio.02366-1412 December 2014 mBio vol. 5no. 6 e02366-14.

123. The Defense Nuclear Facilities Safety Board, an independent organization created by statute in 1988, provides a precedent for regular outside review.

124. A start in this direction for Artificial Intelligence can be found in "Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Food-for-thought paper Submitted by the Chairperson," September 4, 2017, https://www.unog.ch/80256EDD006B8954/(httpAssets)/2117A10B536751D2C1258192004FD7EA/$file/FoodforthoughtPaper_GGELAWS_Final.pdf.

125. Nick Beckstead et al., "Unprecedented Technological Risks" (Future of Humanity Institute, 2014), 10, https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf notes how this technique was used to induce conformance with the nuclear nonproliferation regime and paraphrases an Assistant Secretary General to the U.N., Professor Stephen Stedman, who recommended "an initiative to give developing countries access to safe technologies in exchange for setting up safety and monitoring systems to protect against accidents and terrorism [from biotechnology]."

126. The Nunn-Lugar program usefully has catalyzed such efforts for nuclear weapons. "[R]esponsible systems of governance will need to incorporate clear mechanisms for international dialogue among governing authorities, and perhaps, formal or informal agreements about the use of potential gene drive technologies and comparable stan-

dards for biosafety." National Academy of Sciences, "Gene Drives on the Horizon: Advancing Science, Navigating Uncertainty, and Aligning Research with Public Values," (2016) 163.

127. Our responses to natural events illustrate the need. See the discussion of Fukushima and Ebola crises above, endnote 104.

128. The Department of Defense Global Emerging Infections Surveillance and Response System reflects an acknowledgment of the shared character of health risks and a model for international cooperation. A United Nations Group of Governmental Experts and a 2013 U.S.-Russian agreement envision cyber cooperation against some criminal or terrorist behavior. Thomas Remington et al., "Working Group on the Future of U.S.–Russia Relations: Toward U.S.-Russian Bilateral Cooperation in the Sphere of Cybersecurity," Working Group Paper #7, May 2016, summarizes the United Nations concept: "The practical suggestions of the report include . . . cooperation between states in responding to appropriate requests in mitigating malicious cyberactivity emanating from their territories. For example, in the event of an attack against a U.S. or Russian system originating from a site in the other country, an appropriate response would be for the victim (through its national Computer Emergency Response Team [CERT]) to contact its counterpart in the other country to request a detailed response, for example: '*We believe a cyber attack on System X originated from System Y (IP address: aa.bb.cc.dd), which is within your IP space. Would you please tell us if you have any evidence of this attack that we could analyze together?*' Under the agreement, each side would then exercise the right to narrowly penetrate the ICT space of the presumed attacker to attempt to neutralize the attack. A test of agreement robustness would be whether the two actors jointly conducted the probe (analysis) and neutralization of the cyberthreat."

129. The People's Republic of China, for example, seems more inclined to experiment with gene-editing than present U.S. policy permits, even in the face of acknowledgments that unpredictable "off-target" consequences arise. See David Nield, *Science Alert*, April 11, 2016, http://www.sciencealert.com/scientists-genetically-modify-an-embryo-for-only-the-second-time-ever; "Chinese scientists have genetically modified a human embryo AGAIN: They were trying to genetically modify embryos to be resistant to HIV. While the debate about the ethics of genetically modifying human embryos rages on, scientists in China have successfully carried out the procedure for the second time in history. On this occasion, the team used the CRISPR/Cas9 gene-editing tool to try and create HIV-resistant embryos." This should be seen in the larger context of Chinese attitudes toward eugenics. See Geoffrey Miller, "201: What Should we Be Worried About? Chinese Eugenics," *Edge*, https://www.edge.org/responses/q2013: "China has been running the world's largest and most successful eugenics program for more than thirty years, driving China's ever-faster rise as the global superpower. . . . When I learned about Chinese eugenics this summer, I was astonished that its population policies had received so little attention. China makes no secret of its eugenic ambitions, in either its cultural history or its government policies." Russia and China seem more willing than the United States on moving to rapid automated military systems without human control. More attention and evaluation is needed on these important subjects. Alta Charo of the University of Wisconsin provides a leading discussion of diversity in international approaches to the risks of biotechnology, "Comparative Approaches to Biotechnology Regulation," Presentation at the International Summit on Human Gene Editing, December 1, 2015, https://vimeo.com/album/3703972/video/149182567. As a relatively benign, but illuminating, example of how culture shapes use of technology, anthropologist Jennifer Robertson has suggested: ". . . key cultural factors influence the way in which Japanese perceive robots. First and foremost is Shinto, the native animistic beliefs about life and death. Monotheism has never had a home in Japan, and unlike the three major monotheisms, Shinto lacks complex metaphysical and theological theories and is primarily concerned with notions of purity and pollution. Shinto holds that vital energies or forces called kami are present in all aspects of the world and universe. Some kami are cosmic and others infuse trees, streams, rocks, insects, animals, and humans, as well as human creations, like dolls, cars, and robots." In Jennifer Robertson, "ROBO SAPIENS JAPANICUS: Humanoid Robots and the Post-human Family," *Critical Asian Studies*, 39 no. 3 (2007), 377. See also Aubrey Belford, "That's Not A Droid, That's My Girlfriend," *Countdown*, February 21, 2013, http://www.countdown.org/en/entries/news/s-not-droid-s-my-girlfriend/.

130. These have been buttressed in turn by some instruments like confidence building measures, export controls, classification and sanctions. It is easy to understate or overstate the power of treaties and norms. Robert Keohane provides an insightful discussion, keying on the fact that these understandings establish "practices, regarded as common knowledge in a community, that actors conform to not because they are uniquely best, but because others conform to them as well. . . . [T]hese arrangements . . . establish mutual expectations about others' patterns of behavior and . . . develop working relationships . . . Costs of reneging on commitments are increased, and the costs of operating within these frameworks are reduced." *After Hegemony: Cooperation and Discord in the World Economy* (1984), 89.

131. International Campaign to Abolish Nuclear Weapons, "Nuclear Arsenals," http://www.icanw.org/the-facts/nuclear-arsenals/. A recent meticulous effort at assessment concludes that "the NPT has restrained the spread of nuclear weapons," though very imperfectly so.

132. Matthew Fuhrmann and Yonatan Lupu, "Do Arms Control Treaties Work? Assessing the Effectiveness of the Nuclear Nonproliferation Treaty," *International Studies Quarterly*, September 28, 2016, http://yonatanlupu.com/Fuhrmann%20Lupu%20NPT.pdf.

133. The fact that America refuses to announce its commitment to this norm does not eliminate its force so long as it adheres to it. On the norm of no-first nuclear use see Nina Tannenwald, "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use," *International Organization*, 53 no. 3 (Summer 1999), 433ff; Nina Tannenwald, *The Nuclear Taboo: The United States and the Non-Use of Nuclear Weapons since 1945* (2007); and T. V. Paul, *The Tradition of Non-Use of Nuclear Weapons* (2009).

134. The most substantial quantitative study of the coercive utility of nuclear weapons in international relations concludes: "Coercive nuclear threats have lacked credibility for the first seven decades of the nuclear age in part because of the backlash that would follow any use of nuclear weapons for coercive purposes. . . . [I]f leaders someday decide that his backlash is no longer likely, then the barriers to using nuclear weapons for coercion would begin to weaken." Todd S. Sechser and Matthew Fuhrmann, *Nuclear Weapons and Coercive Diplomacy* (2017), 257.

135. The missile technology control regime (MTCR) created 40 years ago is an example, like the nuclear nonproliferation efforts, of what has and has not been achieved, though through a rather different mechanism. The MTCR is "an informal political understanding among states," now counts 35 "like-minded countries" that "adhere unilaterally," without legal obligation, to common export restrictions. "Frequently Asked Questions (FAQs),"mtcr.info/frequently-asked-questions-faqs/. Arms Control Association, "The Missile Technology Control Regime at a Glance," July 2017, https://www.armscontrol.org/factsheets/mtcr, provides a balanced, brief summary of the missile control regime: "Since its inception, the MTCR has been credited with slowing or stopping several missile programs by making it difficult for prospective buyers to get what they want or stigmatizing certain activities and programs. Argentina, Egypt, and Iraq abandoned their joint Condor II ballistic missile program. Brazil, South Africa, South Korea, and Taiwan also shelved or eliminated missile or space launch vehicle programs. Some Eastern European countries, such as Poland and the Czech Republic, destroyed their ballistic missiles, in part, to better their chances of joining MTCR. The regime has further hampered Libyan and Syrian missile efforts. Yet, the regime has its limitations. Iran, India, North Korea, and Pakistan continue to advance their missile programs. All four countries, with varying degrees of foreign assistance, have deployed medium-range ballistic missiles that can travel more than 1,000 kilometers and are exploring missiles with much greater ranges. India is testing missiles in the intercontinental range. These countries, which are not MTCR members except India, are also becoming sellers rather than simply buyers on the global arms market. North Korea, for example, is viewed as the primary source of ballistic missile proliferation in the world today. Iran has supplied missile production items to Syria." It is notable also that China has announced that it will comply with MTCR guidelines, but its application to join the MTCR group in 2004 was rejected because of its exports of missile technology to North Korea.

136. See U.S. Department of State, "Blinding Laser Weapons (Protocol IV)," https://www.state.gov/documents/organization/190580.pdf.

137. Richard Price, "Reversing the Gun Sights: Transnational Civil Society Targets Land Mines," *International Organization*, 52 no. 3 (Summer 1998), 613–44.

138. United Nations, "Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies," http://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/introouterspacetreaty.html. We have also had success dampening military competition in the Arctic and Antarctic. For a recent discussion of cooperation (and challenges to it) in the Arctic, see Stephanie Pezard et al., "Maintaining Arctic Cooperation with Russia: Planning for Regional Change in the Far North" (RAND, 2018), https://www.rand.org/pubs/research_reports/RR1731.html.

139. Richard Price, *The Chemical Weapons Taboo* (1997).

140. See Peter Katzenstein, ed., *The Culture of National Security: Norms and Identity in World Politics* (1996).

141. The U.S. National Academy of Sciences, Chinese Academy of Science, and the U.K.'s Royal Society collaborated on an International Summit on Human Gene Editing, December 1–3, 2015. http://www.nationalacademies.org/gene-editing/Gene-Edit-Summit/index.htm. The American Biological Safety Association is attempting to articulate norms that can be globally applied. See https://www.absa.org/. Also noteworthy is the "Presidential Commission for the Study of Bioethical Issues, New Directions: The Ethics of Synthetic Biology and Emerging Technologies" (2010).

142. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," July 28, 2015, https://futureoflife.org/open-letter-autonomous-weapons/. Alina Selyukh, "Tech Giants Team Up To Tackle The Ethics Of Artificial Intelligence," National Public Radio, September 28, 2016, http://www.npr.org/sections/alltechconsidered/2016/09/28/495812849/tech-giants-team-up-to-tackle-the-ethics-of-artificial-intelligence, describes a collaboration including Amazon, Facebook, Google, Microsoft, and IBM seeking to drive "research toward technologies that are ethical, secure and reliable – that help rather than hurt – while also helping to diffuse fears and misperceptions about it." Nicholas Bostrom provides an extended discussion of the issues in *Superintelligence: Paths, Dangers, Strategies* (2014).

143. Tim Maurer et al., "Toward A Global Norm Against Manipulating the Integrity of Financial Data," http://carnegieendowment.org/2017/03/27/toward-global-norm-against-manipulating-integrity-of-financial-data-pub-68403. Martha Finnemore and Duncan B. Hollis, "Constructing Norms for GlobalCybersecurity," *American Journal of International Law* (forthcoming) summarizes efforts noting that "[a] UN Group of Governmental

Experts (GGE) and a more inclusive 'London Process' have campaigned for universal cybernorms for all states. Other cybersecurity efforts target norm development for a limited range of actors (for example, like-minded states, major powers) or in specific interest areas (for example, export controls, data protection)." Their footnotes provide detailed references. Note also the NATO effort to develop the "Tallinn Manual on the International Law Applicable to Cyber Warfare,"Nato Cooperative Cyber Defence Centre of Excellence, https://ccdcoe.org/tallinn-manual.html.

144. See Elisa Catalano Ewers et al., "Drone Proliferation: Policy Choices for the Trump Administration" (Center for a New American Security, June 2017), http://drones.cnas.org/wp-content/uploads/2017/06/CNASReport-Drone-Proliferation-Final.pdf; Richard H. Speier et al., "Hypersonic Missile Nonproliferation: Hindering the Speed of a New Class of Weapons" (RAND, 2017).

145. Martha Finnemore and Duncan B. Hollis "Constructing Norms for Global Cybersecurity," *American Journal of International Law*, 2016, provides an excellent overview of how norms may develop. Though Finnemore and Hollis focus on cyber security, their observations are relevant to all technologies. Rebecca Crootof and Frauke Renz, "An Opportunity to Change the Conversation on Autonomous Weapon Systems," https://www.lawfareblog.com/opportunity-change-conversation-autonomous-weapon-systems, argue that a "conversational approach" can influence behavior, sometimes as much as a treaty. Dimitri Kusnezov and Wendell B. Jones, "Are Some Technologies Beyond Regulatory Regimes?" (unpublished draft manuscript, 2017) suggest that rapidly proliferating inexpensive technologies give rise to chaotic situations in which numerous competing participants cannot find stable positions (Nash equilibria). They suggest that "regardless of success in other domains," we will need "novel approaches," perhaps derived from chaos theory, to achieve some stability.

146. The works cited in footnotes in this section reflect a debate among political scientists about the extent to which restraints on international competition are products of inspection by international institutions, individual national abilities to monitor other countries cheating, or internalized taboos. To the degree taboos create some international constraint independent of monitoring, it is hard to believe that important technological activities would remain taboo unless there were reliable reassurance that others were observing the same taboo or at any rate gaining little benefit from any violations.

147. We gain some ephemeral and secondary benefit from restrictions on circulation of technical information, but these are generally weak and especially inadequate when information is relevant to dual use commercial items.

148. The Soviet Union undertook "the construction of the largest biological weapons effort that the world has known to date . . . [at] the very same [time] as Soviet representatives were negotiating the final stages of the BWC [Biological Weapons Convention]." Milton Leitenberg and Raymond A. Zilinskas, *The Soviet Biological Weapons Program: A History* (2012), 9.

149. The same trend is eroding restraints on chemical weapons: "Micro-components require expertise to assemble and have some problems of their own, particularly if their small channels are blocked by impurities. However, they are available for thermal control, mixing, laminar flow, reactor vessels and catalysis and permit reactions at higher temperatures and pressures, instant mixing of reagents and reactants, stable intermediates, and tight control of both exothermic and endothermic reactions. Desktop micro-reactors are now being offered that can provide tons per year of continuous synthetic output by laptop flow chemistry control." See Richard Danzig et al., "Aum Shinrikyo: Insights Into How Terrorists Develop Biological and Chemical Weapons" (Center for a New American Security, 2nd edition, December 2012), fn. 206, 64. Numerous citations with this note provide examples of commercial equipment fitting this description.

150. Elting E. Morison, *Men, Machines, and Modern Times* (1966, 1989), 89.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts and the public with innovative, fact-based research, ideas and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, and was established in February 2007 by co-founders Kurt M. Campbell and Michèle A. Flournoy.

CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and non-partisan. CNAS does not take institutional positions on policy issues. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the authors.

1152 15th Street, NW Suite 950 Washington, DC 20005
t. 202.457.9400 | f. 202.457.9401 | info@cnas.org | cnas.org

Bold. Innovative. Bipartisan.