Governance, Risk, and Artificial Intelligence

Aaron Mannes

■ *Artificial intelligence, whether embodied as robots or Internet of Things, or disembodied as intelligent agents or decision-support systems, can enrich the human experience. It will also fail and cause harms, including physical injury and financial loss as well as more subtle harms such as instantiating human bias or undermining individual dignity. These failures could have a disproportionate impact because strange, new, and unpredictable dangers may lead to public discomfort and rejection of artificial intelligence. Two possible approaches to mitigating these risks are the hard power of regulating artificial intelligence, to ensure it is safe, and the soft power of risk communication, which engages the public and builds trust. These approaches are complementary and both should be implemented as artificial intelligence becomes increasingly prevalent in daily life.*

Able to systematically process, analyze, and transform data at a speed and scale far beyond human capabilities, artificial intelligence (AI) has the potential to augment human health, prosperity, comfort, and knowledge. AI failures, however, could derail this potential. Reaping the benefits of AI, requires managing the risks of AI failures and the use of AI for malevolent purposes. Assessing these risks requires considering not only failures of the AI itself, but how the AI will interact with people and organizations and then identifying approaches to managing these risks.

 ISSN 0738-4602

The following discussion outlines concepts for both hard power and soft power approaches to managing risks associated with AI. Hard power is the power to compel, whereas soft power is the power to persuade and inspire (Nye 2004). Managing the risks of implementing AI will require both approaches. Hard power is the province of formal government regulation and policy. Soft power is the realm of social norms and values. Both hard and soft power approaches to AI risk need to be informed by new research programs. Some of this research will be primarily mathematical. For example, because of the nondeterministic nature of AI, regulatory science will need to develop standards for measuring the risks of AI failure and determining what levels of risk are acceptable. Mathematical approaches are absolutely essential, but not sufficient to manage the risks presented by AI. Mitigating the risks attendant to AI requires understanding how organizations, people, and AI interact. Developing this kind of understanding necessitates interdisciplinary research combining technologists, domain experts, social scientists, and policy experts.

Risk management approaches for AI are needed because when implemented, AI has the potential to fail and cause harms in an array of ways, some unexpected, some spectacular, and some mundane. As one report explained: "… AI systems can suddenly and dramatically fail if the environment or context for their use changes. They can move from super smart to super dumb in an instant." (Scharre and Horowitz 2018). If particularly critical systems (such as components of the power grid or the financial system) are

entrusted to AI without proper safeguards, a failure could have extremely dire consequences. This applies both to truly autonomous systems, as well as to automation systems that place a human-in-the-loop. Ideally, a human in the loop can balance an AI's weaknesses. However, if the human decision-maker does not have a strong understanding of the system's limitations, he or she may exacerbate rather than mitigate the AI's weaknesses. Hopefully risk management programs can limit the super-dumb aspects of AI so that humanity can benefit from super-smart AI.

The analysis that follows does not focus on a particular application of AI, but examines managing the risks of AI from a broader perspective — with a focus on the human aspects of the risk. This discussion is primarily about the civilian sector,[1] examining the AI that is becoming increasingly ubiquitous in daily life around the world. Finally, this discussion considers potential roles of the US federal government, but much of its analysis applies to state and local governments as well as non-US governments and international governmental organizations.

AI Failure: Risk Assessment and Risk Perception

Risk management requires risk assessment. The applications of AI will be varied, so that fully assessing the risks in their deployment in a single article is impossible. This section examines in broad terms the types of risks AI presents: what types of harm might occur when AI fails and how these failures might occur. A fuller consideration of these two issues requires an analysis of risk perception, which will play a powerful role in shaping public reaction to AI failures.

While experts in a system understand risks probabilistically, the general public considers risk through heuristics, which are often informed by sensational stories in the media, personal experience, or the experiences of acquaintances. This gap between expert and layperson can lead to strong and unexpected public reactions to certain types of rare risks and accidents, while giving less consideration to more commonplace risks. Risks that are unknown, poorly understood, unpredictable, and catastrophic inspire feelings of fear and dread. Swimming pools are objectively very deadly, but the risks are known, observable, and limited (Slovic 1987). Terrorism, in contrast, although rare in the United States, is unpredictable and potentially very deadly — and thus inspires feelings of fear and dread, which experts argue is out of proportion to its actual likelihood (Pollack and Wood 2010).

Experts may discount public perceptions of risk, which are not grounded in scientific analysis, but this is unwise. Expert analysis may have its own biases and assumptions, downplaying areas of uncertainty. Public risk perception may have a richer understanding of risk than expert methodology (Slovic 2000). How people feel about AI will shape how it is adopted and deployed. If AI is perceived as harmful, perhaps because of minor incidents that experts discount, there may blow-back from the public.

In this light, the 1978 Three Mile Island nuclear incident should be a cautionary tale for the development and deployment of AI. Although there were no fatalities, and the radiation released had only a minimal impact on public health or the environment (NRC 2018), a strong public reaction effectively froze the nuclear power industry in the United States. The nuclear industry and experts had assured the public that nuclear power was safe, that the public would benefit from inexpensive electricity, and that accidents such as Three Mile Island could not happen. Lending force to the public backlash was the gap between expert and public understanding of the risks of nuclear power. Experts also saw radiation as a well-understood, measurable, and limited hazard. The public, in contrast, saw radiation as poorly understood, dangerous (it was associated with nuclear weapons in the public imagination), and uncontrollable. When the Three Mile Island incident occurred, it became what Slovic (1987) calls a *signal* that actuated public feelings of fear and dread into a backlash against nuclear energy.[2]

AI, which is also poorly understood by most people and is perceived as unpredictable, has a similar potential to spark feelings of fear and dread that could trigger a public backlash. Although they may not have technical merit, popular fears of superintelligent AI that become an existential threat to humanity is an indication of the deep feelings of fear and dread that AI can inspire. When AI fails and causes harm, these feelings could manifest themselves as a signal event.

The types of harms AI failures can cause can be seen on a continuum. At one end of this continuum would be physical harms: property damage, injury, or even death. Autonomous robots, capable of moving and acting in the physical world, are the most obvious, but not the only type of AI that can cause physical harm. It is also possible that decision-support systems can lead to physical damage (as will be discussed later, a human in the loop does not guarantee that AI errors will be spotted). An AI system supporting medical professionals may provide inaccurate diagnoses.

Physical harms are not the only type of harm that AI may cause. AI can cause a range of nonphysical harms, such as financial loss or a privacy violation. In analyzing AI bias, Crawford (2017) identifies harms of allocation and harms of representation. In the case of the former, AI allocates or withholds allocation of resources or opportunities to some groups in favor of others, such as discrimination in granting credit, hiring, or prison sentences. In the case of the latter, harms of representation, AI reinforces discriminatory attitudes and beliefs. Crawford gives many examples, such as AI automatically classifying women as nurses and men as doctors or not processing darker skin tones.

AI can also inflict psychologic harms and harms to dignity. If interacting with AI makes individuals feel alienated or upset, it is harm to individual dignity. AI operating properly, but in a manner that frightens people, such as an autonomous vehicle that drives in ways that people find strange and disturbing,[3] can also do harm. Although these types of harm can be subtle and not as evident as physical injury or financial loss, given the nature of risk perception, it should never be dismissed. Humiliation and fear can provoke deep emotional responses that must be acknowledged and be given full consideration.

Cases of physical harm linked to AI would seem to be the most likely instances to spark a signal event, but the potential of harms to dignity to do so should not be underestimated. Many people already feel themselves at the mercy of vast impersonal forces, particularly when dealing with institutional bureaucracies. In the wake of the Equifax data breach one business columnist captured this frustration, railing against the credit scoring system that "consists of computers from on high influencing our lives via algorithms — without bothering to tell us what's going on." The column goes on to criticize "systems that elevate machines over people" (Sloan 2017). If AI systems exacerbate these feelings, the public reaction may be harsh. The columnist's criticism of algorithms, even though the Equifax data breach was not caused by AI or algorithms, illustrates another important distinction between expert and public approaches to risk. The public may not make fine distinctions between different aspects of technological failures, so that even if the specific failure was in some other part of a system incorporating AI, public perception may lump different types of failures and technologies together.

There are myriad ways in which AI can fail and cause harm. AI will be embedded in complex systems that include not only hardware and software, but people, organizations, and other social entities. Complex interactions between these components can cause an array of failures. Every AI will have unique potential failures, many of which will be driven by the specific environment. The same AI system in two different hospitals, for example, might fail in different ways based on the specific procedures, patients, and personnel of the hospitals. To give an overview of the potential types of failure, two examples will be discussed: self-driving cars and recommendation systems.

The ethical issues around how self-driving cars should respond in emergencies have been a topic of some discussion, but that is only one, very specific way in which a self-driving car can fail. Modern automobiles contain thousands of parts and millions of lines of code. These parts all have the potential to fail, while the software may contain bugs or defects. AI may mitigate some of these failures, while exacerbating others. The suite of sensors and systems to process the sensor data are particularly important elements of an autonomous system. The recent accidents involving self-driving cars appear to involve failures, not of the AI specifically, but rather with either the sensors or the processing of the sensor data (Greenemeier 2018).

Self-driving vehicles will also be interacting with people. This requires operating safely in a complex environment that includes passengers, other vehicles, and pedestrians. Leaving aside the vast potential social implications of self-driving vehicles, interacting with humans raises a number of other complex issues. Criminals and other malicious actors will find ways to use AI. The software in cars will be vulnerable to hacking and the AI may be vulnerable to adversarial machine learning. Autonomous vehicles will also be gathering a great deal of information about their passengers. This raises issues of privacy rights, where even if regulations are adhered to, the feeling that the vehicle is spying on its passengers could lead to public backlash.

AI with humans in the loop is not necessarily less prone to failures, but it changes the types of potential failures. Ideally, systems will be constructed in which the AI and the human in the loop will complement one another's strengths and weaknesses. AI will not suffer fatigue, whereas humans are better able to grasp novel situations and manage ambiguity. At worst people will be incorporated into AI systems, in the words of Madeline Elish (2016), as "moral crumple zones," with little agency in the system. Placing people in this position assigns blame to a human for an AI failure, rather than preventing the failure in the first place. This is a particular issue in dangerous situations, such as motor vehicles or airplanes, where human factors research has shown that people are very bad at instantly turning their attention to a crisis and taking control (Gao 2016).

There is a range of other potential for decision-support systems as well. Many issues with algorithmic bias are the result of decisions about what data are used to train the model. Including and excluding variables, as well as errors in data curation and collection, can skew the AI's results. To take one example, many of the harms of representation described by Crawford (2017) can be traced to training AI on easily available public data sets, which often have biases embedded within them (Levendowski 2018). Harms of allocation can also be tied to problematic data. These cases have drawn particular attention in the criminal justice sphere. Efforts to develop risk scores for probable recidivism have been particularly fraught. In one case, the decision to include past arrests instead of past convictions as a variable can dramatically change how an individual is scored as a risk for recidivism (Wykstra 2018). In another instance, a system that excluded race in its calculations regularly scored African-Americans as being at higher risk for recidivism than comparable whites (Angwin et al 2016).

This process of applying AI to the criminal justice system is being replicated across innumerable fields such as granting loans and mortgages,

personnel decisions, and adjudicating benefits. AI has tremendous potential to serve people and communities with greater speed, efficiency, and equity than human decision-makers (who have biases of their own). There are concerns that AI, trained on existing, flawed data, and overseen by people with little understanding of how AI works, will instead encode and even exacerbate existing biases (Katz 2017).

Even if AI bias is controlled, all AI systems will have limitations. The human decision-maker needs to understand the limits of the system, while the system itself will need to be implemented in a way that meets the needs of the human decision-maker. If a medical diagnosis and treatment system is right most of the time, the medical professionals using it may become complacent and fail to maintain their skills or cease taking pride their work. Further, they might blindly accept decision-support system outputs, not understanding its limitations. This type of failing will be replicated in other spheres, such as criminal justice, where judges have changed their rulings based on risk scores that were later shown to be flawed (Angwin et al. 2016).

Explainable AI can be helpful in managing these risks of AI failing and causing harm. It will be an important component of, but cannot substitute for, a broader, more comprehensive AI risk-management approach. The explanation will be shaped by the nature of the query, which may limit the applicability of the findings, and the findings themselves may be subject to interpretation. Without understanding the broader system within which the AI operates, the explanation may not be meaningful. Finally, for those subject to decisions made by AI, particularly if they have suffered harm, the explanation may not be useful, particularly if they lack a recourse to address any ill effect from the AI decision or action (Ananny and Crawford 2016). Explainable AI will be useful, but it cannot substitute for a broad-based risk management approach to deploying AI.

If AI is to fulfill its promise to improve and enrich lives, the general public will need to trust it. Building this trust requires reducing the possibility and impact of AI failures. As discussed previously, these failures can take many forms, ranging from autonomous vehicles causing traffic accidents, to the Internet of Things allowing criminals new methods to steal or harass, to decision-support systems instantiating biases. Monitoring and testing AI systems so that these many types of failures are limited is a formidable challenge that will require technical knowledge, domain expertise, and careful study of humans and organizations. In the following sections a hard power approach based in government regulation is outlined, followed by a soft power approach of employing risk communication. This is not an either/or proposition. Both approaches have important strengths and complement one another. Both approaches should be considered to address the full range of potential harms that can be caused by AI failures and ensure that AI reaches its potential to increase human health, wealth, and happiness.

Regulation and the Limits of Hard Power

The hard power approach to reducing risk from AI failures is regulation in which governments use rules and directives to ensure that only safe AI systems are deployed. Currently governments regulate a vast range of goods and services to ensure public safety. Because of AI's complexity, its status as an emerging technology, and the broad range of its potential applications, regulating AI will be complicated and there may be limits to its effectiveness. First, regulating AI, including how it will interact with other systems, will be a major technical challenge. Second, while regulatory approaches can be effective at tangible harms, they may be less effective at countering harms of representation or harms to dignity. Third, regulation imposes costs on industries, and poorly designed regulation can harm an industry's growth. Given the enormous potential of AI, poorly designed regulation could represent a suboptimal outcome. This section consists of an overview of the technical and organizational challenges of regulating AI.

At the core of regulating AI is setting standards and issuing guidelines for the safe function of the systems. The strictest regulation is precautionary, which requires systems to meet these standards or not be permitted on the market. Alternately, companies could be granted limited liability if they meet these standards (Scherer 2016), or the standards could be voluntary, but the appropriate agency would provide information about whether the standards were met. An alternative approach is to permit systems to enter the marketplace and monitor their operations and act in the face of failures. The appropriate form of regulation will vary depending on the deployment of AI. Systems that have the highest prospects for catastrophic and deadly failures, such as aircraft, generally undergo rigorous certification before real-world use is permitted. In their recent article, Cummings and Britton (2018) argue that, for AI systems that have the greatest potential to endanger human life and public safety, regulation informed by this strong precautionary approach may be appropriate. It may not, however, be possible to thoroughly test some forms of AI, such as decision-support systems or intelligent agents, without data based on real world operations.

Whatever the enforcement or monitoring mechanisms, actually setting standards for safe and/or effective AI will be a challenge both for regulatory science as well as policy. It will be necessary to develop trials for AI that adequately assess how the system functions, determine the general public's tolerance for risk, and address the knotty ethical issues around testing new technologies that can harm people. Andrew Tutt (2017) makes the useful comparison between pharmaceuticals and nondeterministic

algorithms. He observes that in both cases the results are broadly predictable, but the mechanisms by which the results are achieved are not, and that results and actions in specific instances are not predictable. Regulators will have to define what level of unpredictability is acceptable and then what level and type of testing is appropriate to ensure that this threshold is reached. To take one example, regulators would need to determine how far and in what conditions an autonomous vehicle should be tested. This is complicated by the reality that the process of testing could put people in danger, but without testing in real-world conditions, the system's safety cannot be ascertained. Similarly, a smart medical device, directed by AI, with its potential for emergent behavior, could potentially and unpredictably harm patients, raising difficult ethical considerations.

Developing standards and methods for testing the AI itself is only part of the process. As described earlier, AI will be embedded in a broader system. AI raises complex cyber-security, privacy, and data issues. Robots will be hardware that can be subject to mechanical failures, and all AI systems will interact with people not only as individuals, but also as communities and organizations. In building the capacity to effectively regulate AI, both Tutt (2017) and Scherer (2016) each propose a centralized regulatory authority for AI. Mannes (2016a, 2016b), which considers organizational dynamics and bureaucratic politics, notes that to examine AI in the context of the broader systems in which it operates, domain expertise will be critical. This expertise already exists in current agencies. For these reasons, growing expertise in AI at existing agencies may be the easier path, although this also presents challenges.

Whereas some agencies will have the resources and inclination to take on their new responsibilities to regulate AI, many others will not. Agencies may not be granted the resources needed to regulate an emerging technology. National Highway Traffic Safety Administration, the agency with oversight of autonomous vehicles, has generally taken a limited regulatory approach to this emerging technology, but they have also not been granted any additional resources (Claybrook and Kildare 2018). Other agencies may find themselves regulating AI as a small part of their mission and be unable to devote the resources needed to the mission (Mannes 2016a). Given the huge variety of AI applications and the number of agencies that may become involved in its regulation, there is a danger of a patchwork and inconsistent regulatory approach. There is no simple answer to these problems. Centers of excellence for broad areas of AI applications such as transportation, finance, or criminal justice might facilitate knowledge sharing between agencies and serve as a reservoir of expertise for agencies that are less able to develop it in-house. AI regulatory centers of excellence can also support state and local agencies, which may also lack the resources to fully research these issues themselves.

To take on the challenges of AI governance, agencies will need to recruit the interdisciplinary expertise needed to regulate AI. Technologists, particularly those with expertise in AI, are in short supply and high demand (Shead 2017). The US government struggles to compete to hire and retain this expertise. An important aspect of this challenge is financial, but the US government employs tens of thousands of medical doctors and pays them competitively. Mechanisms exist to make government employment more financially attractive, but it is also important to develop appropriate career paths and offer opportunities that attract talented individuals (Brykcznski et al. 2013). Simply recruiting computer scientists and other specialists in technology is not sufficient. Attorneys and privacy experts with deep understanding of technology will also be needed. Qualitative researchers who can help understand the human aspects of systems in which AI is embedded are also critical. Computational social scientists, who combine technical skill, extensive knowledge of "messy" data (particularly data on people), as well as familiarity with the complex normative issues raised by AI will be particularly valuable (Crawford 2017).

Regulatory approaches to limiting the risks of AI will have important advantages, but also some drawbacks. An effective regulatory regime can reduce the likelihood of the most tangible harms such as physical or financial injury. For AI that can threaten life and limb, direct and robust regulatory regimes are appropriate. Extending a robust regulatory regime to every application of AI may exact costs on a nascent industry and limit the potential benefits. Such formal regulation may not be the most effective means to address harms of representation, harms to dignity, and issues of risk perception. Further, given the complexity of AI itself and the broader systems in which it will be embedded, some failures are inevitable. This is a particular concern as regulators face a learning curve in evaluating this emerging technology. These failures could become signal events and trigger a regulatory blowback. An alternative approach to reducing the likelihood of such failures, while also mitigating the effects of AI failures, is necessary.

Communicating Risk, Building Trust: A Soft Power Alternative

Risk communication can complement the hard power approach of regulation, both by reducing the possibility of AI failure, but also by mitigating the effects of these failures. Risk communication is defined as "the two-way exchange of information, concerns, and preferences about risks between decision-makers and the public" (Portnoy 2010). It is a component of a risk management plan. As will be discussed, when properly conducted, risk communication plays a strategic role in which understanding the intended audience shapes decision-making (Fischhoff 2011).

Understanding how stakeholders perceive risk can shape risk management strategies, including identifying new potential risks. Effective communication also builds trust, and with trust, people will tolerate certain levels of risk if they also understand the benefits. Finally, with this trust, certain types of failure can be better tolerated so that small incidents do not become signal events, and trigger blowback.

Risk communications are often seen as a tactical issue, packaging information. While this is important, it is not a full description of the practice of risk communication. The best-known instances of risk communication are the warnings on pharmaceuticals. These are probably also the least effective form of risk communication. Technical information presented in tiny print on a small folded piece of paper, the typical warning label, in the words of one observer, "Taken as a whole, it fairly shouts: 'Don't read me!'" (Brewer 2011).

These warnings are primarily driven by compliance. The pharmaceutical companies are required to disclose this information and seek to protect themselves from lawsuits, while not giving the patient any incentives not to use their products. The lengthy, jargon-filled, privacy statements accompanying software and social media are similar types of compliance-based risk communication. Many emerging AI developers will probably take a similar approach, but, as discussed earlier, given the potential for AI failure and the ways in which people and society may react to these failures, a more comprehensive approach to risk communication is needed.

There are far more effective means — and significant research devoted — to effectively communicating information that individuals need to weigh risk and benefits. Effectively broadcasting information is important, but not sufficient for effective risk communication about AI. The essence of communication is that it is an interactive and iterative process. A truism in the study of communications is that people overestimate the extent to which they understand the perspectives of other people. Due to this phenomenon, experts may not include information that an audience of lay-people do not know, but would see as essential. At the same time, continuing to focus on information that is known to the audience may reduce their willingness to take in new information (Fischhoff 2011).

Effective risk communication is not an effort by spin-doctors to overwhelm the ill-informed risk perceptions of laypersons with the cold, hard light of reason. Risk communication is an interdisciplinary field incorporating technical fields and the social sciences. Applied risk communication is ultimately about building trust. In the context of risk communication, trust can be understood as reliability and also belief that the other party has one's interests at heart. Where trust exists, individuals and communities will take risks to obtain benefits. Trust requires long-term consistency between words and deeds, and is very difficult to create. It is, however, very easy to destroy with inconsistent actions or statements (Janoske, Liu, and Sheppard 2012). Incidents that are perceived as increasingly risky will have a far greater impact on community perception than incidents that are perceived as indicating safety. In the nuclear power industry, a single accident will dramatically increase public perception of risk, while weeks or months of safe operation will do little to decrease these perceptions. As an exemplar of the importance of trust, medical professionals regularly administer doses of chemicals and radiation that exceed those emitted by nuclear power or chemical industries. The difference is that patients trust that medical professionals have their own interests at heart, and the benefits of the treatment are clearly articulated. The nuclear power and chemical industries have not explained the benefits of their work, and they do not benefit from this trust. Thus, their activities are perceived as riskier (Slovic 2000).

This process of building trust is rooted in two-way communication, where concerns are aired and addressed, with an emphasis on dialogue and consensus. In this process, risk communicators will be learning as much as they are teaching. Risk communication is not a panacea, nor can it be formulaic. It is an expensive, ongoing process. Deploying AI, without public engagement, means the affected public will develop their own narrative, which may cast the AI in a poor light. Then, when a failure occurs, the business or organization that deployed the AI will have little opportunity to gain public trust. Regulation could require that AI firms invest in risk communication, but this approach could result in the industry undertaking risk communication as a matter of compliance. Ideally, those developing and deploying AI should invest in risk communication because it is smart business.

The case for AI risk communication is particularly critical, because AI has the potential to fail in a variety of complex ways, and can provoke feelings of fear and dread; at the same time, AI is becoming increasingly commonplace in day-to-day life. This means AI could spark signal events in which the public perception of risk around the technology increases dramatically. These events may lead to regulatory and public backlash that can stymie the deployment of a technology and potentially deny the public its benefits. Conducted properly, risk communication will engage with the range of stakeholders interacting with AI. By hearing the concerns of these various communities, potential vectors of failure that those creating and deploying AI may not have considered may be revealed and can be addressed. Further, when failures occur, trust will have been built, mitigating the impact. In short, engaging in risk communication can be seen as an effort to fully understand the system in which the AI is engaged. A comprehensive AI risk communication approach that elicits and considers the concerns of stakeholders will result in AI that better serves people and communities.

Risk communication can reduce risk and mitigate each of the sorts of harms that an AI failure can bring.

With a successful risk communication program, people and communities will better understand the AI with which they are interacting, which should reduce accidents and mistakes. Such a program would be an ongoing operation, involving broadcast communications in a variety of formats about the AI, but also meetings with the general public, community leaders, and other critical stakeholders. To be effective, the program would give the stakeholders real input into decision-making. The process of risk communication will also help those producing and deploying AI to understand potential failures in the field and avoid them. Where risk communication can be particularly powerful, and where government regulation may be least effective, is when AI harms individual or community values and dignity. An effective risk communication program can help elicit these kinds of concerns. Most importantly, in building a relationship based on trust, when failures occur — as inevitably they will — stakeholders will be more tolerant and patient.

Ideally, firms providing AI products and services would offer a risk communications program as part of its customer service. If an organization were deploying AI, the vendor would partner with the customer to undertake a risk communications program — identifying stakeholders, conducting outreach, and using the feedback to provide a better service. Finally, risk communication is not only about reaching the public. Internal constituencies and decision-makers also need to be apprised of risks so that they can make informed decisions.

There is a deepening discussion of the ethical, legal, and social implications of AI. A robust risk communications agenda can help instantiate the ethical, legal, and social implication discussions. Effective risk communications require quantitative and qualitative social science, and must address the fundamental questions of values that often underlie perceptions of risk. There is significant overlap between the expertise needed for risk communications and that required for the broader ethical, legal, and social implication dialogue (Mannes 2018).

For much of the technology industry, risk taking is extolled. The mantra is "move fast and break things." The communities that engage these technologies may not share this appetite for risk. Engaging in risk communication to bridge this gap is prudent, and necessary before the wrong thing breaks. If the ultimate goal is for productive and fulfilling human-AI partnerships, it is essential that those building and deploying AI also learn to partner with those who will be using, interacting with, and affected by AI.

Conclusion

AI is a profoundly transformative technology that is fast becoming ubiquitous in everyday life. The rapidity of its spread, tremendous achievements, and enormous potential should not blind enthusiasts to the very real risks from AI, and how the general public may perceive them. This article was written to consider what kinds of risks AI can present and how best to manage and mitigate them. This approach must be holistic and reflect both the many ways in which AI can fail, and the broader system in which a given AI system will operate. There are enormous challenges to ensuring AI is safe, effective, and equitable. Government regulators will need to consider what policy approaches will best foster this emerging technology while also reducing risk. Agencies will also need to develop the necessary technical capabilities to make and implement policies for AI. This hard power approach is important, but not sufficient. As AI becomes more commonplace, and plays a role in people's lives, it can pose risks, not only to safety and security, but also to values and dignity. For these types of risks, a soft power approach of risk communication may be needed. At the core of risk communication is building relationships and understanding. AI is intended to augment and enrich the human experience and enable human endeavors. For this to be achieved, the many stakeholders in the AI project — creators, users, and those interacting with it — must communicate. Significant AI failures are inevitable, but if the stakeholders have relationships of trust, those failures can be minimized and the trust can endure.

Acknowledgments

This article was written with the support of the Data Analytics Technology Center in the Department of Homeland Security Science and Technology Directorate. In no way should anything stated in this paper be construed as representing the official position of the Department of Homeland Security Science and Technology Directorate or any other component of the Department of Homeland Security. Opinions and findings expressed in this article, as well as any errors and omissions, are the responsibility of the author alone.

Notes

1. This article will not discuss the unique and complex issues raised by lethal autonomous weapons systems.

2. Since Three Mile Island, deadly nuclear accidents at Chernobyl and Fukushima have lent credence to public fears, highlighting (Slovic 2000) that public risk perceptions are not merely the product of a lack of expert understanding, but can, at times, be more comprehensive than those of the experts.

3. Autonomous vehicles will be a frequent example, not as a comment on the technology, but rather because they are a readily accessible example.

References

Ananny, M., and Crawford, K. 2016. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic accountability. *New Media & Society* 20(3): 973–89. doi.org/10.1177/1461444816676645.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica May 23. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Brewer, N. 2011. G*oals*. In *Communicating Risks and Benefits: An Evidence-Based User's Guide*. B. Fischhoff, N. Brewer, and J. Downs, editors. Silver Spring, MD: US Food and Drug Administration. 3–10.

Brykcznski, B.; Flattau, P.; and Nek, R. 2013. Attracting and Retaining Science and Engineering Talent in the Federal Government: A Workshop Summary. IDA Document D-4740. Washington, DC: Science & Technology Policy Institute, Institute for Defense Analyses.

Claybrook, J., and Kildare, S. 2018. Autonomous Vehicles: No Driver … No Regulation? *Science* 361(6397): 36–7. doi.org/10.1126/science.aau2715.

Crawford, K. 2017. The Trouble with Bias. Paper presented at the Thirty-First Conference on Neural Information Processing Systems, Long Beach, CA, December 4–9.

Cummings, M., and Britton, D. 2018. Regulating Safety-Critical Autonomous Systems: Past, Present, and Future Perspective. Durham, NC: Duke Human and Autonomy Lab. hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u33/regulating%20autonomy_draft.pdf.

Elish, M. 2016. Moral Crumple Zones: Cautionary Tales in Human Robot Interaction. Paper presented at the 2016 WeRobot Conference. Miami, FL, March 31–April 2. robots.law.miami.edu/2016/wp-content/uploads/2015/07/Elish_cautionary-tales_prelim_draft.pdf.

Fischhoff, B. 2011. Communicating About Analysis. In *Intelligence Analysis: Behavioral and Social Scientific Foundations*. B. Fischhoff, and Chauvin, C., editors. Washington, DC: National Academies Press. 227–48.

Gao, F. 2016. Modeling Human Attention and Performance in Automated Environments with Low Task Loading. PhD dissertation, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA.

Greenemeier, L. 2018. Uber Self-Driving Car Fatality Reveals the Technology's Blind Spots. *Scientific American* 21(March). www.scientificamerican.com/article/uber-self-driving-car-fatality-reveals-the-technologys-blind-spots1/.

Janoske, M.; Liu, B.; and Sheppard, B. 2012. Understanding Risk Communication Best Practices: A Guide for Emergency Managers and Communicator. Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security. College Park, MD: START. www.start.umd.edu/sites/default/files/files/publications/UnderstandingRiskCommunication-Theory.pdf.

Katz, Y. 2017. *Manufacturing an Artificial Intelligence Revolution*. Paper posted November 30, 2017 at SSRN. Amsterdam: Elsevier. papers.ssrn.com/sol3/papers.cfm?abstract_id=3078224.

Levendowski, A. 2018. How Copyright Law Can Fix AI's Implicit Bias Problem. *Washington Law Review (Seattle, Wash.)* 93(2): 579–630. digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1804/93WLR0579.pdf?sequence=1&isAllowed=y.

Mannes, A. 2016a. Institutional Options for Robot Governance. Paper presented at the 2016 WeRobot Conference. Miami, FL, March 31–April 2. robots.law.miami.edu/2016/wp-content/uploads/2015/07/Mannes_RobotGovernance-Final.pdf.

Mannes, A. 2016b. Politics of Robot Governance. Paper presented at the 2016 American Political Science Association Annual Conference. Philadelphia, PA, August 31–September 4.

Mannes, A. 2018. Explaining Autonomy: Risk Communications and Robotics. Paper presented at the 2018 WeRobot Conference. Palo Alto, CA, April 12–14. conferences.law.stanford.edu/werobot/wp-content/uploads/sites/47/2018/02/RobotsRiskComFinal-1.pdf.

Nuclear Regulatory Commission (NRC). 2018. Backgrounder on the Three Mile Island Accident. Washington, DC: U.S. Nuclear Regulatory Commission. www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html.

Nye, J. 2004. *Soft Power: The Means to Success in World Politics*. New York: Public Affairs.

Pollack, J., and Wood, J. 2010. *Enhancing Public Resilience to Mass-Casualty WMD Terrorism in the United States: Definitions, Challenges, and Recommendations*. Washington, DC: Defense Threat Reduction Agency.

Portnoy, P. 2010. Forward. In *Readings in Risk*. T. Glickman and M. Gough, editors. xi. New York: Resources for the Future.

Scharre, P., and Horowitz, M. 2018. *Artificial Intelligence: What Every Policymaker Needs to Know.* Artificial Intelligence and International Security Series. Washington, DC: Center for New American Security.

Scherer, M. 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology* 29(2): 353–400. jolt.law.harvard.edu/assets/articlePDFs/v29/29HarvJLTech353.pdf.

Shead, S. 2017. Tech Giants Are Fighting to Hire the Best AI Talent at the NIPS Conference in LA This Week. *Business Insider* December 4. www.businessinsider.com/tech-giants-are-fighting-to-hire-the-best-ai-talent-at-the-nips-conference-this-week-2017-12.

Sloan, A. 2017. The Equifax Fiasco Is a Classic Case of "Weapon of Math Destruction." The Washington Post September 15. www.washingtonpost.com/business/the-equifax-fiasco-is-a-classic-weapon-of-math-destruction/2017/09/15/ea573004-997f-11e7-b569-3360011663b4_story.html?utm_term=.5b59113bb7d0.

Slovic, P. 1987. Perception of Risk. *Science* 236(4799): 280–5. doi.org/10.1126/science.3563507.

Slovic, P. 2000. Perceived Risk, Trust, and Democracy. In *The Perception of Risk*. P. Slovic, editor. London: Earthscan. 316–326.

Tutt, A. 2017. An FDA for Algorithms. *Administrative Law Review* 69(83). doi.org/10.2139/ssrn.2747994.

Wykstra, S. 2018. Can Racial Bias Ever Be Removed From Criminal Justice Algorithms? Pacific Standard July 12. psmag.com/social-justice/removing-racial-bias-from-the-algorithm.

Aaron Mannes is a senior policy advisor with Culmen International, LLC, where he supports the Department of Homeland Security Science & Technology Directorate's Data Analytics Technology Center. Mannes has previously worked at the University of Maryland Institute for Advanced Computer Studies modeling terrorism and international security affairs. He earned his doctorate in policy studies from the University of Maryland College Park and has written several books and numerous articles on international affairs and technology policy.