RAND CORPORATION

Perspective

Expert insights on a timely policy issue

The Risks of Artificial Intelligence to Security and the Future of Work

Osonde A. Osoba, William Welser IV

Introduction

Overview

Future generations may look back at our time and identify it as one of intense change. In a few short decades, we have morphed from a machine-based society to an information-based society, and as this Information Age continues to mature, society has been forced to develop a new and intimate familiarity with data-driven and algorithmic systems. We use the term *artificial agents* to refer to devices and decisionmaking aids that rely on automated, data-driven, or algorithmic learning procedures (including artificial intelligence (AI) in its many manifestations).[1] These include devices as banal as Roomba robots and online recommendation engines to more advanced cognitive systems like IBM's Watson. Such agents are becoming an intrinsic part of our regular decisionmaking processes. Their emergence and adoption lead to a bevy of related policy questions. How do we reorient our thinking on relevant policy in this new regime? Where are our blind spots in this space? How do users, as well as affected populations, identify and remedy errors in logic or assumptions? What sectors are the ripest for disruption by artificial agents, and what approaches to regulation will be most effective?

We wrote a previous report (Osoba and Welser, 2017) emphasizing the existence of blind spots and bias with respect to artificial agents in the criminal justice system, but other sectors will likely be impacted. This Perspective discusses the outcome of a structured exercise to understand what other areas might be affected by increasing deployment of artificial agents. We relied on a diverse group of experts to paint scenarios in which AI could have a significant impact. The Research Methodology section describes how we did this elicitation, fundamentally an exercise in forecasting. Thus, not

all elicited domains or scenarios will be important (either in terms of likelihood of occurrence or magnitude of impact). Therefore, we focus this discussion on two high-value domains out of the full set of highlighted domains: (1) security and (2) future of work. We conclude with a section on policy themes and suggestions for how to approach AI-related issues and concerns.

Preamble and Context

The maturation of the Information Age has forced some adaptation and evolution in our laws, regulations, and policies. But the pace and intensity of technological change has often made it difficult for the policy, regulations, and laws to keep up. As has been the case in other periods of intense change, the lag in the evolution of laws and regulations can lead to significant policy gaps.

For example, data-laden societies are currently re-evaluating acceptable personal standards of privacy. This is necessary given the growing use of ubiquitous data collection and powerful, cheaply run, and readily available algorithms. The legal standards of reasonable or acceptable privacy need renegotiation to accommodate new technologies that are being adopted at pace and scale. There is a lot at stake (Ohm, 2009; Davis and Osoba, 2016): health data privacy, consumer fairness, and even the constitutional Census mandate.

There is also a re-evaluation of our legal understanding of innovation. The U.S. Supreme Court's *Alice Corp. Pty. LTD. v. CLS Bank International* (2014) decision shows the law still adapting to delimit the extent of patentable innovation in an information-based society (McKinney, 2015). The case tackles the question of what qualifies as tangible patentable innovations. The patent system originated in an environment in which valuable innovations were tangible physical innovations. And the patent system favored the tangible implementations of such physical innovations. Valuable innovations in the Information Age (e.g., Google's PageRank algorithm) are increasingly better characterized as intangible ideas (or "abstract ideas"): business methods, algorithms, or procedures. Do we recognize only tangible physical novelty as innovations? Or do innovative-yet-intangible algorithms or methods count as protected innovation? Patent law has changed significantly over the past 30 years to try to accommodate such new types of innovation. Laws will likely continue to adapt. At stake are incentives for commercial innovation and clear property rights.

The legal standards of reasonable or acceptable privacy need renegotiation to accommodate new technologies that are being adopted at pace and scale.

With the advent of widespread automation and AI, employment is another area seemingly primed for significant upheaval. The common fear is that automation and AI will displace human workers in the labor market, leading to rising and runaway unemployment. Older expressions of this fear focused on *automation* as the antagonist. AI is essentially automation with the added capacity for learning or adaptation. Thus, AI represents a natural generalization of automation.

Examples like Blockbuster stoke the fire for such mass unemployment fears. Before its demise, Blockbuster employed more than 60,000 workers. Netflix, as of late 2016, leveraged AI (enabled via

machine learning) technology and existing information infrastructures to not only serve but also enhance the value delivered to consumers with just about 3,500 workers. Technology giants like Google and Facebook are deriving outsized value from the AI research and development segments of their workforces.[2] AI R&D is driving much of their value to users, although their AI workers within commercial companies account for less than 10 percent of their workforce. On a larger scale, Frey and Osborne (2013) present a controversial analysis claiming that about 47 percent of current U.S. workers are in occupations that are at risk of displacement to automation over the next two decades.

This Perspective represents an attempt to highlight potential policy challenges ahead as AI becomes more central in the private, commercial, and public spheres. We approach the topic in a careful manner using available current literature to bolster our discussion. The next section describes our methodology for structuring our inquiry.

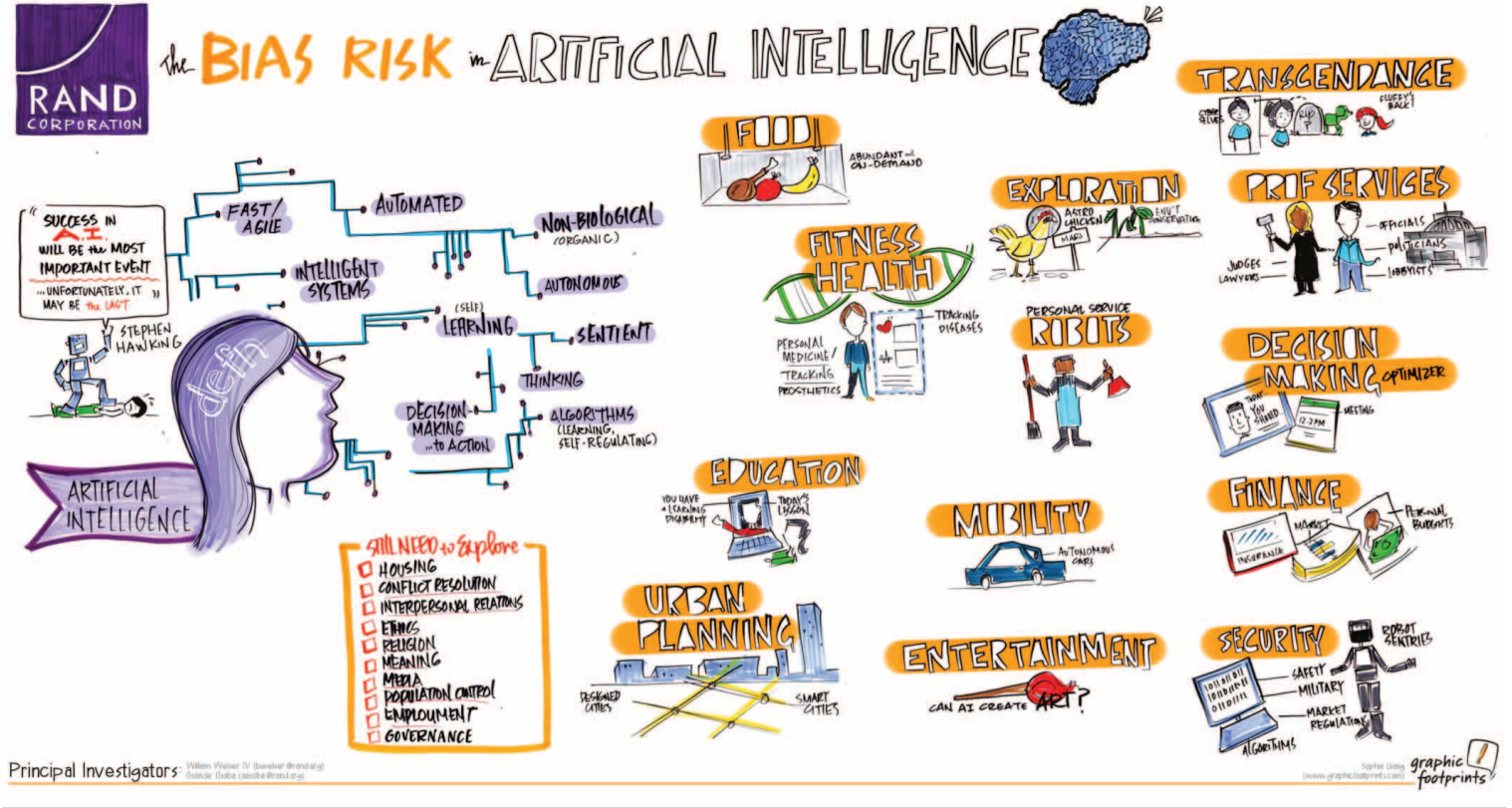
Research Methodology: An Interdisciplinary Approach to AI Research

Our discussion so far may seem to foreshadow impending instability because of AI. Popular discussion on AI and algorithms tends to share a similar tone. We proposed to try to cut through the hype with analytic and cross-disciplinary thinking on the risks and future of AI. We convened a team of RAND researchers from across the academic disciplines and with a multitude of professional experiences to discuss AI. We curated a team of colleagues who were diverse in gender, ethnicity, and race while also making sure that we did not overrepresent for deep technical knowledge of AI. The team included expertise in economics, psychology, political science, engineering, mathematics, neuroscience, anthropology, and design. Our hope was that, by convening such a group of researchers with extensive and varied training, we would encourage a dialogue around AI that was distinct and would allow for insights from topics and substance adjacent to AI.

The group's first exercise was to take part in a structured brainstorming session involving independent thought, small group discussions, and whole-group debate to first develop a working definition of AI and then to highlight application areas most prone to disruption by AI. The working definition, represented on the left of Figure 1, was developed via a rapid-fire collection of answers to "describe AI in less than five words." The themes of the contributions are summarized in the visual in Figure 1 (the rendition samples some defining notions of AI and highlights some areas susceptible to change due to AI or automation). The group condensed these further into a working definition for AI: It is an

"autonomous, non-biological learning system." Then, we asked that colleagues provide ideas of both near-term and future applications for AI for effecting individuals as well as society. The proposed applications were binned into themes that are depicted on the right side of Figure 1. In particular, our team of colleagues pointed to areas like decisionmaking, security, and even "transcendence,"[3] as application areas ripe for future AI innovation. After discussing the opportunities and risks related to these applications, the group suggested themes that were not covered by the previous discussion, but where AI is just as relevant. Those themes are captured under the "Still Need to Explore" box and included critical areas of governance, conflict resolution, and media.

Figure 1. Artist's Rendition of Outcome of RAND Panel Exercise



SOURCE: Sophia Liang, www.graphicfootprints.com.
NOTE: This piece samples some defining notions of AI and highlights some areas susceptible to change due to AI or automation.

We followed these initial exercises by driving the team to deeper discussion via a future-casting exercise[4] for which the larger group was split into four subgroups. While the insights and outcomes of this part of the activity varied depending on the envisioned future that the groups chose, there was a list of application spaces that were included by each of the teams. Due to the consistency across groups, we consider those to be "no-brainer" applications of AI, and they include

- security (national and domestic)
- employment ("future of work")
- decisionmaking
- health.

Following the activities with our team of colleagues, we chose to dive more deeply into the literature on the first two topics, security and employment, with a goal of developing a clearer picture of the risks inherent in the use of algorithms or artificial intelligence (or jointly as artificial agents) in these spaces. We chose these topics because we believe they are more pressing concerns to governments and the populace. To discuss AI benefits and risks as related to the security of nation states, we convened a small team of colleagues with deep knowledge of political science and defense analysis. The team also highlighted other key areas as potential game-changing applications of AI: conflict resolution/dispute mediation, advanced surveillance, and cybersecurity. We do not explore these in this piece.

Security

We can divide our security discussion in two parts: national security and domestic security. We use *national security* as an umbrella term to discuss risks that external state and nonstate actors pose to a country. We use the term *domestic security* as an umbrella term to stability risks that emerge from within the nation state.

The application of AI to surveillance or cybersecurity for national security opens a new attack vector based on this data diet vulnerability. Adversaries may learn how to systematically feed disinformation to AI surveillance systems, essentially creating an unwitting automated double agent.

National Security

Our first discussion of AI-related risks in national security brought up familiar themes. For example, fully automated decisionmaking in the national security space can lead to costly errors and fatalities. Cold War anecdotes (and movie plots) abound about countries brought to the brink of nuclear war by malfunctioning automated nuclear defense systems. A recent DefenseOne piece (Lohn, Parasiliti, and Welser, 2016) by RAND researchers looks at the thorny question of AI weapons without human mediation.

Cybersecurity was identified as a particularly fertile area for AI-enabled vulnerabilities. A key function of artificial agents (both informational and cyberphysical artificial agents) is the efficient manipulation of information. Thus, artificial agents may be particularly suited to information warfare and cybersecurity applications.

Augmenting Internet of Things (IoT)–targeting malware like Mirai[5] (Newman, 2017) with intelligence can vastly improve the strategic potential of malware. Stuxnet[6] (Langner, 2011) is an illustrative example of how decisive, advanced, and strategically targeted malware can be. One factor restricting intelligence in malware is the need for small malware payloads to prevent detection. For example, the payload for the intelligent malware, Stuxnet, was larger than most malware (Zetter, 2010). But it is conceivable that future developments in swarm or distributed AI may result in strategic botnets with small malware payloads but devastating effects.

We also identified another familiar and important concern. There is a *data diet vulnerability* found in much of current autonomous learning systems (Osoba and Welser, 2017). AI systems are typically only as good as the data on which they are trained. They crystallize any biases or falsehoods found in their training data (Barocas and Selbst, 2016). The application of AI to surveillance or cybersecurity for national security opens a new attack vector based on this data diet vulnerability. Adversaries may learn how to systematically feed disinformation to AI surveillance systems, essentially creating an unwitting automated double agent. Recent work already demonstrates the viability of such training set poisoning attacks for machine-learning–based malware detection systems (Biggio et al., 2012; Biggio et al., 2014; Huang et al., 2017). The corruption of Microsoft's AI chat-bot (Lee, 2016) provides a more popular demonstration of this line of attack on a commercial AI system. This could be an opportunity for counterintelligence operations.

There is another interesting blind-spot (or feature) for security in an AI-enabled world: the use of network intervention methods by foreign-deployed AI. American intelligence agencies have reported that they believe the recent 2016 U.S. election cycle was subject to undue foreign interference via foreign cyberattacks (Paletta, 2016). These attacks presented as selective public releases of hacked private data in an attempt to affect voters' opinions. While this type of attack is detectable and recognizable given the right information, more advanced artificial agents could make malevolent actors more effective and less detectable in this process.

For other related blind spots, consider, for example, the implications of growing personalization in Internet services. Heavy personalization in Internet use can lead to the creation of personalized *filter bubbles* (Pariser, 2011). This can have the effect of sharply segregating political discourse across groups. The hypersegmentation of the information-consuming population provides opportunities for highly targeted political messaging. This targeting ability can lead to a reduced emphasis on truth in messaging and journalism; our susceptibility to confirmation bias makes us more likely to believe messages that confirm pre-existing beliefs, however false. Tufekci (2016a) argues that this demographic hypersegmentation, our cognitive biases, and the closed nature of our online social media platforms (Tufekci, 2016b) results in echo chambers that amplify misinformation.

Observers and academic researchers have argued that both major parties benefited from or took advantage of the ability to target (sometimes fake) news and messages during the 2016 election cycle (Allcott and Gentzkow, 2017; Love and Cooke, 2016). Subsequent elections and referenda now have to contend with this sort of influence campaigns (Byrne, 2016). There is no indication that artificial agents were applied with intent to exploit this vulnerability even though news curation algorithms are part of what makes this vulnerability possible (Dewey, 2016). But future artificial agents feeding our information consumption habits could be trained to make

more strategic use of this vulnerability or alternatively nullify it via purposeful, systematic injection of noise or disinformation (part of what has been more recently referred to as "fake news").

Alternatively, an artificial agent with a comprehensive view of political and social networks (information that is increasingly easy to collect) may be able to spot network influence opportunities to achieve a political outcome. The agent may be able to intervene in networks to connect isolated but similar-minded groups that, with increased scale and geographical diversity, can act to achieve specified political outcomes. Such activity would constitute a larger scale and more strategic version of current advanced targeting of political messages on social media, e.g., using Facebook's ad-targeting platform. Recent conversations in the emerging field of information operations are beginning to highlight other automatable practices for enabling targeted influence campaigns (Waltzman, 2017).

Domestic Security

Our second discussion of the use of artificial agents in domestic security highlighted significant AI-related risks. A visceral example of such risks is the deployment of artificial agents for the surveillance of civilians by governments. Oliver Stone's recent movie "Snowden" discusses one such example. Government surveillance, at best, speaks to a government's intent to act. Intent may not carry the same moral or legal weight as actions themselves. But this distinction can be a harder sell when the government in question is repressive. Less trusting appraisals of U.S. government surveillance (Alexander, 2012) argue that surveillance in the United States has not been a neutral tool historically. Inequitable surveillance, however legal prima facie, can be a tool for entrenching inequity. The increasing sophistication of artificial agents enables surveillance by all resourceful governments—repressive and benevolent alike—and in the extreme case invokes thoughts of George Orwell's *1984*.

Expanded search and seizure capabilities available to law enforcement organizations and the attendant erosions of privacy are the most obvious concern.

The use of artificial agents in the domestic security space is already common and is thus not mere speculation. Legal scholars (Citron, 2007) have written extensively on the use of algorithmic or data-driven systems for surveillance and for administrative law (e.g., in the administration of welfare benefits). Previous RAND reports (Perry et al., 2013) discuss the use and limits of predictive policing algorithms in U.S. civilian law enforcement. ProPublica's recent report (Angwin, Larson, Mattu, and Kirchner, 2016) on machine bias describes the use of algorithms in the criminal justice proceedings. The report discusses the use of a recidivism estimation algorithm, COMPAS, in the criminal justice system for parole hearings. The COMPAS system was shown to give systematically biased results. This bias, compounded by the misguided use of the system for bail and sentencing proceedings, led to significant inequities in criminal sentencing outcomes in the courts utilizing the technology.

The growing use of artificial agents in law enforcement can trigger concerns about fundamental citizens' rights. Expanded

search and seizure capabilities available to law enforcement organizations and the attendant erosions of privacy are the most obvious concern. Another manifestation of this concern crops up in response to the use of traffic cameras. These devices may be minimally intelligent, but there were early concerns about potential 6th Amendment rights violations inherent in the use of evidence produced by nonhuman automated agents. These concerns have since been dismissed in some jurisdictions (e.g., within the state of California subsequent to *People v. Goldsmith* (2014), which sets California precedent in affirming that evidence from red light cameras is not hearsay when authenticated by an officer's testimony). Another recent manifestation is in the use of robots in the apprehension of law-breakers. A recent U.S. mass shooting event ended with the shooter's death at the hands of a robot-delivered bomb (Murphy, 2016). Some observers expressed discomfort at this new development in law enforcement, and what it might mean for the presumption of innocence.

The natural question at this stage is: How fundamental is the tension between citizens' legal rights and AI? Given that police departments and law enforcement professionals have gone to great lengths to gain and build trust with their communities, is there a chance that unanticipated errors in automation could needlessly undercut these efforts? We will have to negotiate around such questions as we adapt the law to our new capabilities.

Future interplay between legal rights and artificial agency will continue to be an area of concern. Legal scholars are beginning to explore other implications related to legal personhood for artificial agents (Bayern, 2015; LoPucki, 2017). Bayern describes how artificial agents can gain legal personhood. Corporate charter competition among states has made it easy for artificial agents to easily gain and maintain legal personhood via corporate charter. LoPucki discusses how such algorithmic legal entities enjoy a comparative advantage over human-controlled entities in criminal, terrorist, or other antisocial activities partly due to jurisdiction and the ease of software transfer across borders. This could potentially be novel legal terrain.

Artificial agents are now increasingly able to do a growing share of tasks that we have typically relied on humans to do through the labor market. This includes medical/radiological diagnosis, driving vehicles, writing specific types of news reports, and other tasks.

Future of Work

Questions around the future of work tend to crop up in discussions on AI. By future of work, we mean the effect of AI on the supply of and demand for human labor. The key focus of anxiety in this space is the extent to which advances in AI enable artificial agents to do tasks cheaply and thereby replace human agents who earn income by doing those tasks. In an older discussion on the topic, Moravec (1998) depicts the concern viscerally (although possibly inaccurately). He imagines tasks as lying in a field of plains, hills, and mountains in which the (objectively measured) cognitive

difficulty of a task is reflected by its altitude in that landscape. At the mountaintops, we have tasks like social interaction, hand-eye coordination, locomotion, etc. At various hilltops, we have tasks like playing chess or Go, image recognition, and others. Advancing AI is akin to a rising flood in this landscape. Over time, AI systems will grow to be competent at many lowland and hilltop tasks. And we will be left with the mountaintops. The key future-of-work question becomes: What are those mountaintops?

Inquiry into AI and future of work extends a long tradition of research into the effects of automation on the labor market (e.g., Armer, 1966; Karoly and Panis, 2004; Autor, 2015; Acemoglu and Restrepo, 2017). Automation concerns have traditionally been more focused on robotic systems with limited independent intelligence or adaptability (e.g., industrial robots, ATMs) usually applied to relatively lower skilled tasks (Autor, 2015; Acemoglu and Restrepo, 2017). These systems act based on explicitly programmed instructions. Newer AI systems do not need such explicit guidance and can act based on insights learned from data or experts.[7] Artificial agents are now increasingly able to do a growing share of tasks that we have typically relied on humans to do through the labor market. This includes medical/radiological diagnosis, driving vehicles, writing specific types of news reports, and other tasks. Much of the research on automation effects still applies. The question now becomes: How does the new capacity for automation-by-AI change the labor picture?

An effective labor market serves at least two key purposes: to provide labor to do productive work and to provide a source of workers' income through earned wages. The future of work question considers what types of effects (and how much) the rise of advanced artificial agents has on the effective operation of the labor market, especially to the ability of workers to earn meaningful wages. This is an emotionally and politically charged consideration given the economic, social, and cultural functions that employment serves in human societies. The uncertain recovery of the global economy since the crash of 2008 also served to stoke employment fears, especially given the aspects of the recovery that has led it to be typified as a "jobless recovery"—one where macroeconomic growth occurs despite employment levels that either stay the same or fall. This trend also coincides with another recent trend showing a steady decline in the labor share of national income compared to the share of national income going to capital (Karabarbounis and Neiman, 2014; Baker, 2016; Autor et al., 2017a). Therefore, labor (compared with capital) is getting a lower share of returns from economic growth and increased national productivity.

The net effect is a labor market that is increasingly weaker at funding workers' standards of living even as automated systems (including AI systems) do a growing share of total productive work. Some tech executives, economists, and policy analysts (Reeves, 2016; Murray, 2016; *The Economist*, 2016) have responded to these concerns by calling for decoupling wages/living standards from employment via universal basic income (UBI) or guaranteed income schemes. There are cost, incentives, and administration difficulties with these schemes. Small government pilots already exist or are planned (Marica in Brazil and Alaska in the United States). But the jury is still out on UBI feasibility at large scales and over long terms.

Near- to Medium-Term Trends

Near-term trends in our new AI-enabled world show that artificial agents are having a disruptive effect on traditional work patterns (*The Economist*, 2014). This disruption is not always negative. Disruption has resulted in new labor opportunities. Irani, for example, writes

about the emergence and features of "microwork" (Irani, 2015). Microwork refers to short-term tasks or jobs for humans like survey-taking, driving, cleaning, and other tasks. Historically, the cost of crowdsourcing and flexibly coordinating employment contracts for these types of work has been prohibitive. AI-powering platforms like TaskRabbit, Uber, Lyft, and Amazon's Mechanical Turk have reduced these costs dramatically. And microwork-coordinating services have grown in response. Concerns remain about the status and benefits of microworkers in the labor markets (Cherry, 2015).

Economists (Jaimovich and Siu, 2012; Autor, 2015) also write about observed differential susceptibility of jobs to automation and loss in recent economic data. They find that middle-skill routine-based jobs (e.g., production, manufacturing, operators) have been historically more susceptible to higher job losses during economic busts and slower recovery in economic booms than low-skill (e.g., janitors) and high-skill (e.g., software engineers) jobs. They call this effect "job polarization." Other near-term labor market effects of artificial agents include a deskilling effect by which automation leads to the loss of specialized human abilities or skills. Automation reduces labor demand for people with the skills in question. And workers reorient away from learning skills that have already been automated during their training. But the advent of division of labor had a similar deskilling effect. Therefore, the importance of this effect is still uncertain. A potential positive effect of AI deployment includes improved commuting because of safer autonomous unmanned vehicles (Anderson et al., 2016).

One concern with the growth of AI systems is that the investment required for AI development is available only to a very restricted few, such as very-high-tech firms, firms with access to large databases, and highly skilled technical workers. This means that

Automation reduces labor demand for people with the skills in question. And workers reorient away from learning skills that have already been automated during their training.

returns and productivity gains from automation-by-AI accrue to a very restricted group of "superstar firms" (Autor et al., 2017b). At the same time, if jobs continue to be automated, then the basic income-generating function of labor diminishes. This has the effect of further increasing income inequality at the national and global levels.

Frey and Osborne (2013) attempt to answer the question of the susceptibility of jobs to automation. Their study hypothesizes a plausible set of factors that determine an occupation's susceptibility to automation: the occupation's requirement for creativity intelligence, social intelligence, and fine perception and manipulation. Then they rank U.S. occupations based on these three factors, and they use this approach to estimate that 47 percent of U.S. workers are at high risk of displacement by automation. A study by the Organisation for Economic Co-operation and Development (OECD) (Arntz, Gregory, and Zierahn, 2016) challenges this 47-percent estimate. Arntz et al. argue that it is tasks in a job that are subject to automation; jobs just morph to include both automated and human tasks. This task-based approach to occupations is a more detailed, and potentially more accurate, analysis of the evolution of occupations. And it allows for a more careful accounting of the interactions between tasks and skills within occupations

(Autor, 2013). The OECD study finds that only 9 percent of jobs in 21 OECD countries are at risk for full automation under this task-based analysis. In many cases, when certain tasks are automated, other tasks are added to a worker's repertoire, and overall value is enhanced. Another report by the International Labor Office asserts that susceptibility analyses are only appropriate at the economic-sector level (Chang, Rynhart, and Huynh, 2016).

A Framework for Examining Occupational Susceptibility to Automation

In thinking about the long-term trend, it is worth pointing out a couple of trends in this area of research. First, our historical track record on forecasting susceptible jobs has been pretty abysmal. Brynjolfsson and McAfee (2014) discuss stark examples of failures in this forecasting space. For example, Queen Elizabeth famously refused to grant a machine patent in the 16th century out of concern for the jobs of her subjects. Nevertheless, England went on to usher in the Industrial Revolution, during which even children were used to satisfy the intense demand for labor. Levy and Murnane argued, circa 2004, for the infeasibility of autonomous unmanned vehicles (Levy and Murnane, 2004). Recent developments at Uber, Google, and Tesla (as well as conventional vehicle manufacturers) show this forecast to be off the mark.

We can contrast these failed predictions with Paul Armer's work. Armer wrote a RAND report (Armer, 1966) for the U.S. National Commission on Technology, Automation, and Economic Progress in the late 1960s. His analysis was prescient about the implications of an information-based society. He predicted current predicaments like our renegotiation of privacy, the value of information, the effect of information technology on the pace of technological advancement, and other scenarios. His predictions on employment were in much broader strokes. For example, he predicted the need for more time spent in education and the necessity of continuous workforce retraining to match the pace of technology changes. His more specific discussions (e.g., on the use of Internet and communication technology in banking, education, and sales) seem to have been borne out.

Another point worth considering: We also have not always done a stellar job forecasting what tasks are difficult for artificial agents to learn. Moravec's paradox (Rotenberg, 2013; Moravec, 1998) describes this shortcoming: We judge tasks as computationally difficult when they require significant human focus (such as proving theorems, playing chess, or playing Go) and yet undervalue legitimately hard computational tasks that seem to require less human effort (such as perception, creativity, social interaction, hand-eye coordination). In other words, we are poor at objective estimations of cognitive and processing difficulty. This bias in judgment makes forecasting the evolution of work error-prone.

The recurrence of future of work concerns suggests that researchers have not found a good frame for judging the susceptibility of occupations and tasks to automation—more specifically, automation by artificial agents. AI researchers and economists have begun to articulate the limit of AI automatability (Ng, 2016; Autor, 2015): Essentially, AI excels at tasks that are well-defined, repetitive or routine, and for which performance is easy to judge. But these boundaries are still malleable and subject to new research. Minsky's early definition of AI (Minsky, 1961) identified planning as a key AI subdomain. The current difficulty AI systems experience with chaotic environments reflects a slight lag in AI progress in the subfield of automated planning (Geist, 2017).

In general, susceptible occupations are those that are composed of automatable tasks that interact in simple, well-defined ways. This mirrors the current restriction of artificial agents to well-defined problems. With this in mind, we propose a framework building on RAND research on managing occupational surprise (Baiocchi and Fox, 2013). Our framework suggests that two factors determine an occupation's susceptibility to automation:

1. *Amount of chaos* that a worker must contend with regularly in the occupation. This loosely refers to how many tasks or scenarios the worker must learn to manage individually and how often the worker needs to switch between scenarios. The complexity of the scenarios and the diversity of skills required to manage the scenarios also play a role. Highly chaotic occupations include Firefighting, Navy SEALs, Politicians, and Surgeons.
2. *Typical response times* required for a worker to perform tasks in the occupation effectively.

Figure 2 depicts this framework. We posit that this framework allows us to distinguish between the susceptibility of various types of occupations to automation. This framework is something of a refinement of the routine versus cognitive occupation framework used in Jaimovich and Siu, 2014, based on work by Autor, Acemoglu, and others. It also mirrors Autor's (2015) discussion on what it takes to make tasks easy to automate with AI or machine-learning. Specifically, we suggest that occupations that fall in the low-chaos environments are more amenable to automation by AI, with longer-response-time occupations being the most amenable. High-chaos occupations are harder to automate. And occupations characterized by long-response times are more amenable than those that feature short-response times.

The effect of typical response times on susceptibility to automation will be modulated by chaos or amount of task-switching required. That modulation may be complex. Long response times and low levels of chaos indicate longer periods spent on single tasks with limited switching between different types of tasks. The automated planning problem (automated determination of how and when to switch among which tasks) in this regime is relatively easier to solve. Therefore, occupations in these regimes are likely to have higher susceptibility to automation. Automated planning in the high-chaos/short-response-time regime is a much harder problem. High chaos means there is a larger number of distinct functions an AI system needs to be designed to accomplish along with being able to respond to the randomness of the order of presentation of the tasks being accomplished—designing an AI to perform acceptably in such a regime involves planning for and managing significant complexity mixed with very short timeframes.

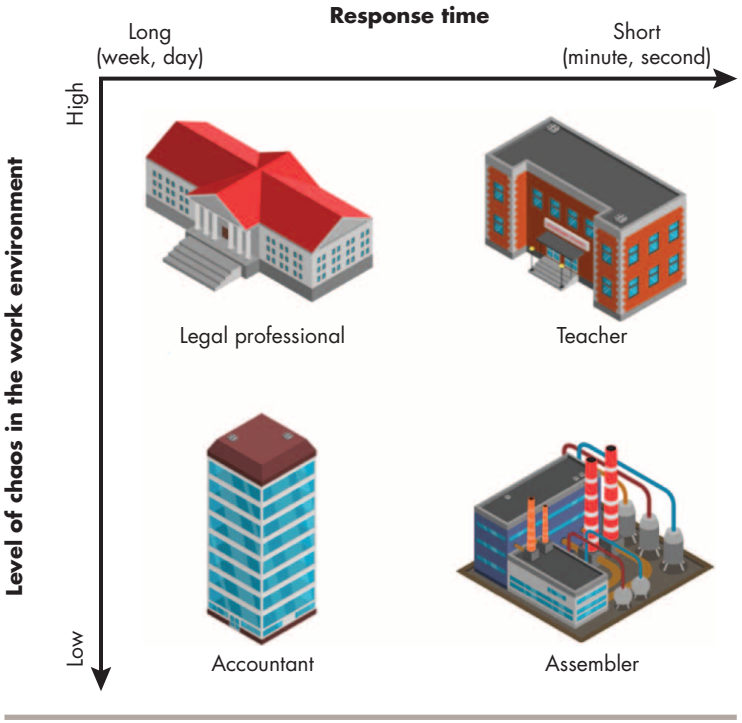
The low-chaos/short-response-time has historically been a regime of high automation activity (e.g., assembly line settings). In the low-chaos/short-response-time case, the short response times required can be physically taxing for human worker, but since there is a smaller repertoire of tasks, the design and planning challenges are of much lower complexity. These challenges can be adequately addressed through things such as careful factory layout and portfolios of statically-programmed robots. Basic automation, not necessarily automation-by-AI, has sufficiently handled this regime along with the maturation of the modern factory (e.g., the automobile industry from the late 1950s to current day).

The high-chaos/long-response-time represents an interesting automation paradox. In this regime, there are still a lot of functions to master and a significant automated planning problem but at

longer time-scale. Many of these functions require a fine-tuned ability to navigate (often unspoken) social and cultural norms. It is not yet clear if more intelligent AI will be capable of making the fine-grained integrated decisions required to excel in this regime even if there is no time crunch.

While this framework should provide useful guidance to designers and developers of artificial agents, it also indicates that many of the automation efforts to date may not necessarily have targeted the full swath of occupations amenable to automation.

Figure 2. Suggested Framework for Characterizing Occupational Susceptibility to Automation



Thus far, the majority of automation efforts have been in factory-like environments (possibly for safety reasons). This discrepancy appears to indicate that automation has disproportionately affected occupations that require less formal education, like factory workers, instead of more white-collar occupations, like accountants. This is arguably because of the lower cognitive requirements and higher mechanical aspects for that quadrant, leading to a more straightforward approach toward robotic design. But advances in AI are raising the bar on the cognitive capacities that automation-via-AI can exhibit (Moravec's metaphorical rising flood). We suggest that this framework could help identify the next most realizable opportunities for automation-via-AI.

General Themes Identified and Suggestions

Our discussions and survey of AI impacts in the security domains and employment domains highlighted some broader risks and themes. We will try to capture these highlights here. It may be possible to combat AI risks individually as they arise, but it is likely that a more holistic approach will provide higher leverage.

We can identify general themes from our discussions. The first theme focuses on a key difference in the attention frames for human- versus AI-information processing. One of the reasons the filter bubble phenomenon exists is because we seek to avoid information overload. Or in the decisionmaking frame, we delegate away subsidiary information-processing tasks to allow us to focus on the key decisions. Our human brains currently have *limited flexibility* that contributes to limited attention frames. Artificial agents' attention frames can be *more flexible*, and the available scope often improves with new innovations in information technology.[8] "Big data" (now including data streams from the IoT)

expands the limits of what artificial agents *can* pay attention to. Specific AI implementations may use restricted attention frames. But the state of technology allows for massive expansion of those frames compared to what humans can accommodate. This difference in scopes of attention may underlie many of our blind-spots in assessing AI impacts.

The second theme some of these discussions highlight is a type of *diminished resilience* due to limited information on how automation influences our lives. By this we mean the potential to reduce systemic resilience or increase systemic fragility by relying more on artificial agents. The Flash Crash of 2010 is an example of the introduction of new modes of system fragility due to the use of artificial agents (Nuti et al., 2011). This diminished resilience often takes the form of an overly trusting or insufficiently critical view of artificial agents. This unwarranted trust is already evident in our dealings with the rudimentary artificial agents we currently use. It is a manifestation of the human tendency toward *automation bias* (Osoba and Welser, 2017). Our previous discussion of automation bias (Osoba and Welser, 2017) highlighted the documented human tendency to ascribe more credibility to outcomes and decisions produced by artificial agents without accounting for the error and bias risks inherent in these agents. Pariser's discussion on filter bubbles and other recent revelations on Facebook's news curation algorithms highlight how easy it is to obscure the role of algorithms in daily life. That role will, without doubt, grow over time. This can have significant systemic effects (e.g., the filter bubble phenomenon, hyperpolarization of online discourse). Automation bias also has significant implications for accountability in decisionmaking (e.g., questions of appeal in the criminal justice system).

Many automated systems are not able to recognize when they are in error states, especially when these error states relate to social norms.

One reactionary approach to combating such effects would be to deploy additional layers of artificial agents to correct gaps or deficiencies with the existing agents. But the task of automatically correcting bad AI outcomes or stringing together multiple AI implementations can be difficult. And both approaches are often more difficult than just designing an agent to automatically recognize bad outcomes. Many automated systems are not able to recognize when they are in error states, especially when these error states relate to social norms. In general, an effective automated AI regulator may need to be as complex as the system. And, in theory, such regulators would also require regulation.

A more natural minimal response would be to require disclosure-style transparency. The focus here would be to highlight domains where artificial agents act autonomously or mediate access to information. This focus may help analysts identify which system's behavior to audit or, at the very least, foster more critical attitudes toward artificial agents. Less minimally, the difficulty of automated AI regulation makes a limited argument in favor of human-in-the-loop regulation of automated systems, especially for critical systems or systems with high-assurance requirements.

We can try to crystallize these themes more formally and make suggestions to address them. These are not meant to be exhaustive.

Theme 1: Artificial Agents Are Fundamentally Attention-Multipliers

This highlights a key difference in the attention frames for human- versus AI-information processing. We highlighted filter bubbles above as an example of the consequences of our limited attention frames. Artificial agents' expanded attention frames (e.g., enabled via big data and IoT streams) can be more flexible, and their scope often improves with new innovations in information technology. This difference in scopes of attention may underlie many of our blind spots in assessing AI risk.

Suggested response: Get ahead of the potential negative impacts of information and arbitrage discovery. Government needs to have access to AI expertise and insights. Decisionmakers must be as proactive as possible in detecting and addressing new vulnerabilities AI presents to risk-sensitive areas like security. It is also equally important to try to identify new positive affordances that AI presents. This requires creativity and cross-pollination between government and sectors of AI excellence.

Theme 2: Reliance on Artificial Agents Increases the Risk of Diminished Resilience

The second theme highlighted is a type of diminished resilience due to limited information on how automation influences our lives. This diminished resilience takes the form of the deskilling effect in the employment discussion. It is easy for important skills (e.g., making fire) to be lost once the skill has been automated. Diminished resilience can also take the form of an overly trusting or insufficiently critical view of artificial agents. This unwarranted trust is already evident in our dealings with the rudimentary artificial agents we currently use. It is a manifestation of the human tendency toward automation bias. The role of AI in all aspects of 21st-century life will only grow. This *will* have significant systemic effects.

Suggested response: Develop systematic procedures for enumerating dependences on artificial agents and consider appropriate less automated failsafe procedures. It may not be possible to rely on further automation and AI to address problems caused by AI. And such a reliance may not be scalable and risks exacerbating the resilience problem. As a first step, we can emphasize the development of more intelligent approaches for human audits of automated systems (informational AI and cyberphysical systems). This includes encouraging disclosure-style transparency and formalizing AI behavioral validation routines. Also encourage the use of human-in-the-loop regulation for select high-assurance or high-accountability applications of artificial agents. More work needs to be done to explore AI fail-safes in more depth.

Theme 3: AI Has the Potential to Cause Rapid Economic and Social Disruption

The impact of AI has arguably been muted in the past. That muted impact led to what were called "AI winters," in which stakeholders (business, government, people) lost faith in the promise of AI. This round of AI hype has been more productive. Commercial investments (from Google, Amazon, Uber, and others) in ripe technologies (deep learning, recommendation systems, unmanned control, and more) have led to impactful breakthroughs (autonomous cars, natural language processing, automatic language translation, etc.). The speed and scope of socioeconomic effects due to AI could be significant and even unprecedented. There are already significant

employment and regulatory impacts with the rise of workers in the gig economy, which is enabled by AI platforms. An insufficient response to AI socioeconomic impacts can inequitably disenfranchise significant portions of the population (e.g., via displaced jobs) and pose risks to national stability. There is more work to be done to better evaluate and forecast these impacts. More generally, there is a case to be made for designing robust adaptive regulatory schemes to match the pace of technological progress.

Suggested response: Recognizing that this is a potentially novel socioeconomic regime, decisionmakers must adjust evaluations of policy risk accordingly. This includes being more open to evaluating nonstandard intervention options (e.g., other safety net schemes like basic income). Part of the concern is that policymakers are generally slow and reactive to changes in technology generated by commercial entities. Additionally, the speed of impact makes a reactive stance potentially costlier.

Theme 4: AI Has Geopolitical Implications

The United States enjoyed substantial advantages in innovation and economic strength in the second half of the last century. The reasons for this state of affairs and the prospects of its continuation are secondary considerations.[9] The point to focus on is that many of the innovations in AI occurred in the United States and the attending benefits of these innovations accrued to the United States *first*. These innovations are diffusing quickly, especially with the strong academic and commercial push to "democratize AI." The rise of AI expertise and innovators in other nations (e.g., China's Baidu, Alibaba, and Didi) is probably the more indicative signal pointing to the loss of the United States' first-mover's advantage in the AI space. Further complicating the arena is the fact that the United States has ceded dominance in high-performance computing as those assets have proliferated and become accessible globally. It is no longer tenable to assume the absence of foreign actors with comparable AI expertise and resources.

Suggested response: Assumptions of enduring superiority in AI-related technology and expertise should be discarded to account for the reality that stiff global competition now exists. The competition may grow stiffer over time as the quality of math and science education in the United States (as measured in cross-national education surveys like the OECD's PISA [undated]) continues to rate as "average" or "below average." Decisionmakers could adopt a "race-to-the-moon" stance with an appropriately aggressive strategy to invest in AI-related research and infrastructure. The AI talent pool is and will continue to be a space of tight geopolitical competition. The standard lever would be to improve STEM educational outcomes the local K–12 pipeline. This is admittedly a complex task that will likely only yield results at longer time-scales. The AI talent concern serves to make this task more urgent. There is also significant strategic importance to attracting and securing such talent, and interested nations should acknowledge that the pool is global in nature and thus requires immigration policies that prioritize these skill sets. For the United States, immigration policy is a major lever for three reasons: (1) an uncharacteristically high percentage of U.S.-resident AI experts are foreign-born or first-generation immigrants; (2) post-graduate programs where AI expertise develops have been relying on student immigration for many years now. The U.S. K–undergraduate education pipeline has not been supplying enough native graduates interested in supplying enough

native graduates interested in STEM; and (3) the global competition for experts may be close to a zero-sum game. It is easier to cultivate and retain U.S. tech dominance if experts and would-be experts immigrate from other competing states.

Conclusion

This Perspective explores the implications of AI prevalence on two key policy-relevant areas: security and employment. Our focus was on highlighting the potential vulnerabilities and inequities that the use of AI imposes on these two dimensions of society. A team of RAND colleagues with diverse expertise and experiences identified these two areas among others as deserving of careful attention in the age of AI. Other areas identified include (not explored in this piece): health, decisionmaking (broadly), conflict resolution/dispute mediation, and cybersecurity. The cross-disciplinary nature of the problems we found illustrates the need to continue to engage researchers and analysts with a diverse set of expertise and experiences in order to inform policy decisionmakers as to positions and actions to pursue with respect to artificial agents, and more broadly, AI.

In our exploration of the current and potential future effects of AI on security and the future of work, we identified the following crosscutting themes in AI impacts.

1. Artificial agents are fundamentally attention-multipliers and can have unexpected and serious systemic effects.
2. Reliance on artificial agents increases the risk of diminished resilience.
3. AI has the potential to cause unprecedented rapid economic and social disruption.
4. The employment and migratory preferences of the global AI R&D talent pool are questions of significant geopolitical concern.

We use these themes to help identify concrete suggestions on how researchers and policymakers can better respond to the policy implications of AI.

Notes

[1] There have been numerous attempts to define AI canonically. McCarthy (2007) defines intelligence as "the computational part of the ability to achieve goals in the world." Minsky (1961) gave an enumeration of functions required to achieve artificial forms of intelligence: search, pattern recognition, learning, planning, and induction (or generalization from observed examples). Any artificial system or process performing any of these functions (broadly construed) to achieve goals in the world will qualify as AI for the purpose of our discussion.

[2] Value in terms of the extent to which AI or machine learning (ML) permeates their business operations compared with the size of their AI research divisions. The relative sizes of their overall divisions to their AI research workforce can serve as a rough proxy. The pure AI R&D divisions in these companies account for less than one-tenth of their workforces as of late 2016.

[3] By *transcendence*, participants presumably meant the capability to exceed physical human limitations by transferring minds to silicon. We had encouraged our panelists to be open-minded. But for our research purposes, we chose not to pursue that line of inquiry much further.

[4] *Future casting*, in this sense, describes a type of world-building activity that we used to think through various future states of the world. The group was split into four smaller groups and asked to consider one quadrant of a two-by-two matrix that allowed for four distinct future states of the world. Group members were challenged to consider the potential role of AI in their respective future state and to project the areas for application with the most promise, and those with the most risk or downside.

[5] Mirai capitalizes on the relative insecurity of newer IoT devices and marshals large populations of these ubiquitous network-enabled devices into coordinated botnets to execute massive distributed denial of service (DDoS) attacks of unprecedented capacity. The Mirai botnet was responsible for the massive network outage on October 21, 2016, that disrupted large-scale operations like Twitter, GitHub, and Netflix.

[6] Stuxnet is a malware that infects SCADA systems commonly used as controllers in industrial systems. Stuxnet was allegedly responsible for the destruction of Iran's industrial nuclear centrifuges.

[7] Such flexible artificial agents circumvent Polanyi's paradox (Autor, 2015) that states that human expertise consists of more than we can tell or teach. AI systems are increasingly capable of learning desired expertise as long as there are examples (data) from which to learn—even if we cannot explicitly articulate the desired expertise.

[8] The key counterlever to AI's more flexible attention-frame is computational efficiency. An expansion to accommodate a larger available scope of attention can easily make a task computationally infeasible without careful design. A reviewer highlighted the recent research activity on effective attention mechanisms for AI systems.

[9] Goldin and Katz (2009) make the case that shortcomings in the U.S. educational system make the continuation of U.S. economic exceptionalism unlikely.

References

Acemoglu, Daron, and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets," National Bureau of Economic Research, NBER Working Paper No. 23285, 2017. As of October 11, 2017: http://www.nber.org/papers/w23285

Alexander, Michelle, *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*, New York: The New Press, 2012.

Alice Corp. Pty. Ltd. v. CLS Bank International, 573 U.S. ___, 134 S. Ct. 2347, June 19, 2014. As of October 11, 2017: https://www.supremecourt.gov/opinions/13pdf/13-298_7lh8.pdf

Allcott, Hunt, and Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*. National Bureau of Economic Research, NBER Working Paper No. w23089, 2017.

Anderson, James M., Nidhi Kalra, Karlyn Stanley, Paul Sorensen, Constantine Samaras, and Tobi A. Oluwatola, *Autonomous Vehicle Technology: A Guide for Policymakers*, Santa Monica, Calif.: RAND Corporation, RR-443-2-RC, 2016. As of October 11, 2017: https://www.rand.org/pubs/research_reports/RR443-2.html

Angwin, Julia L., Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks," ProPublica, 2016. As of December 7, 2016.

Armer, Paul, *Computer Aspects of Technological Change, Automation, and Economic Progress*, Santa Monica, Calif.: RAND Corporation, P-3478, 1966. As of December 8, 2016: http://www.rand.org/pubs/papers/P3478.html

Arntz, M., T. Gregory, and U. Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis," *OECD Social, Employment and Migration Working Papers*, No. 189, 2016. As of December 8, 2016: http://dx.doi.org/10.1787/5jlz9h56dvq7-en

Autor, David H., "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *The Journal of Economic Perspectives*, Vol. 29, No. 3, 2015, pp. 3–30.

Autor, David H., "The 'Task Approach' to Labor Markets: An Overview," *Journal for Labour Market Research*, Vol. 46, No. 3, 2013, pp. 185–199.

Autor, David H., David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "Concentrating on the Fall of the Labor Share," *American Economic Review, Papers and Proceedings*, Vol. 107, No. 5, pp. 180–185, May 2017a.

Autor, David H., David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," CEPR Discussion Paper No. DP12041, May 2017b. As of October 11, 2017: https://ssrn.com/abstract=2968382

Baiocchi, Dave, and D. Steven Fox, *Surprise! From CEOs to Navy SEALs: How a Select Group of Professionals Prepare for and Respond to the Unexpected*, Santa Monica, Calif.: RAND Corporation, RR-341-NRO, 2013. As of November 16, 2016: http://www.rand.org/pubs/research_reports/RR341.html

Baker, Brian J., "The Laboring Labor Share of Income: The 'Miracle' Ends," *Monthly Labor Review*, U.S. Bureau of Labor Statistics, 2016. As of November 16, 2016: http://www.bls.gov/opub/mlr/2016/beyond-bls/the-laboring-labor-share-of-income-the-miracle-ends.htm

Barocas, S., and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, Vol. 104, 2016.

Bayern, Shawn J., "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems," *Stanford Technology Law Review*, Vol. 19, No. 93, October 31, 2015. As of October 11, 2017.

Biggio, Battista, Blaine Nelson, and Pavel Laskov, "Poisoning Attacks Against Support Vector Machines," *Proceedings of the 29th International Conference on Machine Learning*, Cornell University, 2012. As of October 11, 2017: https://arxiv.org/abs/1206.6389v1

Biggio, Battista, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli, "Poisoning Behavioral Malware Clustering," *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, 2014, pp. 27–36.

Brynjolfsson, E., and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies,* New York: W. W. Norton & Company, 2014.

Byrne, Andrew, "Macedonia's Fake News Industry Sets Sights on Europe," *Financial Times*, December 15, 2016. As of October 12, 2017: https://www.ft.com/content/333fe6bc-c1ea-11e6-81c2-f57d90f6741a

Chang, Jae Hee, Gary Rynhart, and Phu Huynh, *ASEAN in Transformation: How Technology is Changing Jobs and Enterprises (Working Paper No. 10)*, Switzerland: Bureau for Employers' Activities, International Labour Office, 2016. As of December 8, 2016.

Cherry, M. A., "Beyond Misclassification: The Digital Transformation of Work," *Comparative Labor Law Journal & Policy Journal*, Vol. 37, 2015, p. 577.

Citron, D. K., "Technological Due Process," *Washington University Law Review*, Vol. 85, No. 6, 2007, p. 1249.

Davis, John S., and Osonde A. Osoba, *Privacy Preservation in the Age of Big Data: A Survey*, Santa Monica, Calif.: RAND Corporation, WR-1161, 2016. As of September 12, 2017: https://www.rand.org/pubs/working_papers/WR1161.html

Dewey, Caitlin, "Facebook Has Repeatedly Trended Fake News Since Firing its Human Editors," *Washington Post*, October 12, 2016. As of October 11, 2017: https://www.washingtonpost.com/news/the-intersect/wp/2016/10/12/facebook-has-repeatedly-trended-fake-news-since-firing-its-human-editors/?utm_term=.6c3ecb67d0d7

Frey, C. B., and M. A. Osborne, "The Future of Employment: How Susceptible are Jobs to Computerisation?" *Oxford Martin Programme on Technology and Employment*, September 7, 2013. As of December 7, 2016.

Geist, Edward Moore, "(Automated) Planning for Tomorrow: Will Artificial Intelligence Get Smarter?" *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017, pp. 80–85.

Goldin, Claudia Dale, and Lawrence F. Katz, *The Race Between Education and Technology*, Cambridge, Mass.: Harvard University Press, 2009.

Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, "Adversarial Attacks on Neural Network Policies," Cornell University, 2017.

Irani, L., "The Cultural Work of Microwork," *New Media & Society*, Vol. 17, No. 5, 2015, pp. 720–739.

Jaimovich, Nir, and Henry E. Siu, *The Trend is the Cycle: Job Polarization and Jobless Recoveries (Working Paper 18334)*, Cambridge, Mass.: National Bureau of Economic Research, 2012.

Karabarbounis, Loukas, and Brent Neiman, "The Global Decline of the Labor Share," *The Quarterly Journal of Economics*, Vol. 129, No. 1, 2014, pp. 61–103.

Karoly, Lynn A., and Constantijn (Stan) Panis, *The 21st Century at Work: Forces Shaping the Future Workforce and Workplace in the United States*, Santa Monica, Calif.: RAND Corporation, MG-164-DOL, 2004. As of October 11, 2017: https://www.rand.org/pubs/monographs/MG164.html

Langner, Ralph, "Stuxnet: Dissecting a Cyberwarfare Weapon," *IEEE Security & Privacy*, Vol. 9, No. 3, 2011, pp. 49–51.

Lee, P., "Learning from Tay's Introduction," Microsoft, March 25, 2016. As of December 7, 2016: http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Levy, Frank, and Richard J. Murnane, *The New Division of Labor: How Computers Are Creating the Next Job Market*, Princeton, N. J.: Princeton University Press, 2012.

Lohn, A., A. Parasiliti, and W. Welser, "Should We Fear an AI Arms Race? Five Reasons the Benefits of Defense-Related Artificial Intelligence Research Outweigh the Risks—for Now," Defense One, February 8, 2016. As of December 7, 2016: http://www.defenseone.com/ideas/2016/02/should-we-fear-ai-arms-race/125670/

LoPucki, Lynn M., "Algorithmic Entities," 95 *Washington University Law Review*, UCLA School of Law, Law-Econ Research Paper No. 17-09, April 17, 2017. As of October 12, 2017: https://ssrn.com/abstract=2954173

Love, Julia, and Kristina Cooke, "Google, Facebook Move to Restrict Ads on Fake News Sites," Reuters, 2016. As of November 16, 2016: http://www.reuters.com/article/us-alphabet-advertising-idUSKBN1392MM

McCarthy, J., "What is Artificial Intelligence?" Stanford University, November 12, 2007. As of October 12, 2017: http://www-formal.stanford.edu/jmc/whatisai.pdf

McKinney, David, "Alice: Tumbling Down the Rabbit Hole of Software Patent Eligibility," *UMKC Law Review*, Vol. 84, 2015, p. 261.

Minsky, M., "Steps Towards Artificial Intelligence," *Proceeding of the IRE*, January 1961, pp. 8–18.

Moravec, Hans, "When Will Computer Hardware Match the Human Brain," *Journal of Evolution and Technology*, Vol. 1, No. 1, 1998, p. 10.

Murphy, M., "The Dallas Police Department Used a Bomb Robot to Take Out Last Night's Sniper," Quartz, 2016. As of November 16, 2016: http://qz.com/727153/the-dallas-police-department-used-a-bomb-robot-to-take-out-last-nights-sniper/

Murray, C., "A Guaranteed Income for Every American; Replacing the Welfare State with an Annual Grant is the Best Way to Cope with a Radically Changing U.S. Jobs Market—and to Revitalize America's Civic Culture," *Wall Street Journal*, June 3, 2016. As of November 16, 2016: http://www.wsj.com/articles/a-guaranteed-income-for-every-american-1464969586

Newman, Lily H., "The Botnet That Broke the Internet Isn't Going Away," WIRED, December 9, 2016. As of November 7, 2017: https://www.wired.com/2016/12/botnet-broke-internet-isnt-going-away/

Ng, Andrew, "What AI Can and Can't Do," *Harvard Business Review*, November 9, 2016. As of November 16, 2016: https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

Nuti, Giuseppe, Mahnoosh Mirghaemi, Philip Treleaven, and Chaiyakorn Yingsaeree. "Algorithmic Trading," *Computer*, Vol. 44, No. 11, 2011, pp. 61–69.

Ohm, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review*, Vol. 57, 2009, pp. 1701–1777.

Organisation for Economic Co-operation and Development, "Programme for International Student Assessment (PISA)," web page, undated. As of October 11, 2017: http://www.oecd.org/pisa/

Osoba, Osonde A., and William Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*, Santa Monica, Calif.: RAND Corporation, RR-1744-RC, 2017. As of September 12, 2017: https://www.rand.org/pubs/research_reports/RR1744.html

Paletta, D., "U.S. Blames Russia for Recent Hacks; Intelligence Agencies Believe Hacks are Meant to 'Interfere with the U.S. Election Process,'" *Wall Street Journal*, October 7, 2016. As of December 7, 2016: http://www.wsj.com/articles/u-s-blames-russia-for-recent-hacks-1475870371

Pariser, Eli, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, United Kingdom: Penguin Books, 2011.

People v. Goldsmith, 326 P.3d 239, 59 Cal. 4th 258, 172 Cal. Rptr. 3d 637, June 5, 2014.

Perry, Walter L., Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*, Santa Monica, Calif.: RAND Corporation, RR-233, 2013. As of December 7, 2016: http://www.rand.org/pubs/research_reports/RR233.html

Reeves, Richard V., "Time to Take Basic Income Seriously," Brookings, 2016. As of November 16, 2016: https://www.brookings.edu/opinions/time-to-take-basic-income-seriously/

Rotenberg, V., "Moravec's Paradox: Consideration in the Context of Two Brain Hemisphere Functions," *ANS: The Journal for Neurocognitive Research*, Vol. 55, No. 3, 2013.

"Sighing for Paradise to Come," *The Economist*, June 4, 2016. As of November 16, 2016: http://www.economist.com/news/briefing/21699910-arguments-state-stipend-payable-all-citizens-are-being-heard-more-widely-sighing

"There's an App for That," *The Economist*, December 30, 2014. As of November 16, 2016: http://www.economist.com/news/briefing/21637355-freelance-workers-available-moments-notice-will-reshape-nature-companies-and

Tufekci, Z., "Mark Zuckerberg is in Denial," *New York Times*, November 15, 2016a. As of November 16, 2016: http://www.nytimes.com/2016/11/15/opinion/mark-zuckerberg-is-in-denial.html

Tufekci, Z., "As the Pirates Become CEOs: The Closing of the Open Internet," *Daedalus,* Vol. 145, No. 1, 2016b, pp. 65–78.

Waltzman, Rand, *The Weaponization of Information: The Need for Cognitive Security*, Santa Monica, Calif.: RAND Corporation, CT-473, 2017. As of September 12, 2017: https://www.rand.org/pubs/testimonies/CT473.html

Zetter, Kim, "Blockbuster Worm Aimed for Infrastructure, But No Proof Iran Nukes Were Target," WIRED, September 23, 2010. As of October 11, 2017: https://www.wired.com/2010/09/stuxnet-2/

About This Perspective

The growth of artificial intelligence (AI) and algorithmic systems in society and government presents new risks. The broad applicability of AI systems means that a wide swath of domains will be affected and are potentially susceptible to new and unexpected failure modes. The set of affected domains include health, education, security, employment, and finance, to name a few.

This Perspective piece lays out the risks in two domains of significant importance and public interest: security and employment. These domains are only a subselection of a larger set of affected domains identified by a panel of experts. We drill down on the near-to-medium–term trends and implications of AI proliferation in these domains. In brief, we highlight the potential for significant disruption because of AI proliferation in the subdomains of cybersecurity, governance, justice (criminal and civil), and labor market patterns. Our discussion of the future of work also presents a novel framework for thinking about the susceptibility of occupations to automation. The Perspective ends with a set of AI policy recommendations informed by the trends we highlight.

This Perspective and the recommendations should be of interest to decisionmakers, especially those tasked with managing security or labor. This Perspective is also intended to highlight the sort of multidomain analysis needed to address AI policy implications.

The authors are grateful for the sponsorship provided by RAND's Center for Global Risk and Security (CGRS) and RAND Ventures. We are also very grateful for our reviewers' valuable advice on making this work better. We would like to thank Ian Goodfellow and John Davis for their comments on the whole piece. We would also like to thank Lynn Karoly for her insightful comments on the Future of Work section in this Perspective.

About the RAND Center for Global Risk and Security

This work was conducted within the RAND Center for Global Risk and Security (CGRS). CGRS works across the RAND Corporation to develop multidisciplinary research and policy analysis dealing with systemic risks to global security. The Center draws on RAND's expertise to complement and expand RAND research in many fields, including security, economics, health, and technology. A board of distinguished business leaders, philanthropists, and former policymakers advises and supports the center activities, which are increasingly focused on global security trends and the impact of disruptive technologies on risk and security. For more information about the RAND Center for Global Risk and Security, visit www.rand.org/international/cgrs.

RAND Ventures

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier, and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND Ventures is a vehicle for investing in policy solutions. Philanthropic contributions support our ability to take the long view, tackle tough and often controversial topics, and share our findings in innovative and compelling ways.

RAND's research findings and recommendations are based on data and evidence, and therefore do not necessarily reflect the policy preferences or interests of its clients, donors, or supporters.

Funding for this venture was provided by gifts from RAND supporters and income from operations.

About the Authors

Osonde A. Osoba is an engineer at RAND and a professor at the Pardee RAND Graduate School. His research work has focused on the application and policy implications of algorithms and artificial intelligence.

William Welser IV is a senior management scientist and the director of the engineering and applied sciences department at RAND. His research work has focused on emerging technology and science and technology policy more generally.

For more information on this publication, visit www.rand.org/t/PE237.

www.rand.org

PE-237-RC (2017)