by: Edward Parson, Richard Re, Alicia Solow-Niederman and Elana Zeide

Introduction: AI Advances, Impacts, and Governance Concerns

Artificial intelligence (AI), particularly various methods of machine learning (ML), has achieved landmark advances over the past few years in applications as diverse as playing complex games, language processing, speech recognition and synthesis, image identification, and facial recognition. These breakthroughs have brought a surge of popular, journalistic, and policy attention to the field, including both excitement about anticipated advances and the benefits they promise, and concern about societal impacts and risks – potentially arising through whatever combination of accident, malicious or reckless use, or just social and political disruption from the scale and rapidity of change.

Potential impacts of AI range from the immediate and particular to the vast and transformative. While technical and scholarly commentary on AI impacts mainly concerns near-term advances and concerns, popular accounts are dominated by vivid scenarios of existential threats to human survival or autonomy, often inspired by fictional accounts of AI that has progressed to general super-intelligence, independent volition, or some other landmark similar to or far surpassing human capabilities. Expert opinions about the likelihood and timing of such extreme further advances vary widely.[1] Yet it is also increasingly clear that advances like these are not necessary for transformative impacts – for good or ill, or more likely for good *and* ill – including the prospect of severe societal disruption and threats.

The potential societal impacts of AI, and their associated governance challenges, are in significant ways novel. Yet they also lie in the context of prior concerns with

by: Edward Parson, Richard Re, Alicia Solow-Niederman and Elana Zeide

assessing and managing technology-related risks, which has been an active area of research and policy debate since controversies over technologies related to energy, environment, weapons, computation, and molecular biology in the 1960s and 1970s.[2] This work has generated many insights into societal impacts and control of technology, of which two in particular stand out. First, the societal impacts of technology are not intrinsic to the technology, but emerge from the social processes by which technologies are developed and applied. It is thus not possible to assess or manage societal impacts by examining a technology divorced from its economic, political, and social context.[3] Second, these linked pathways of technological and social development are complex and uncertain, so societal impacts cannot be confidently projected in advance, no matter how obvious they may appear in retrospect. There is thus a structural tension that hinders efforts to manage the impacts of technology for social benefit. The process of developing, applying, and reacting to any new technology both gradually clarifies its effects, and also builds constituencies with interests in its unhindered continuance and expansion. Efforts to manage impacts thus move from an early state in which they are limited in knowledge because the nature of impacts is not clear, to a later state in which impacts are clearer but politically difficult to manage.[4] This paradox is not absolute or categorical, but does describe a real tension and a real challenge to effective assessment and management of technological risks – which is clearly applicable to efforts to manage AI risks today.[5]

While every technological area is in some respects unique, AI is likely to be more challenging in its potential societal effects and governance needs than even other contentious, high-stakes, rapidly-developing technologies. There are at least three reasons for this, rooted in characteristics of AI and applications that are likely to be

enduring. First, AI is weakly defined. It includes diverse methods and techniques, derived from multiple prior areas of inquiry, which have fuzzy boundaries with each other and with multiple other areas of technological advance and challenge. The weak definition and boundaries of AI make it difficult to precisely localize and specify related objects of concern and their boundaries, and thus difficult to define regulatory authority or other forms of governance response. Second, AI has a foundational character. AI advances promise to transform, and interact strongly with, other areas of technological advance such as computational biology, neuroscience, and others. AI's foundational role suggests comparison to historical examples of transformative technologies on the scale of those that drove the industrial revolution – electricity, the steam engine and fossil fuels. Considered together, the diffuse boundaries and foundational character of AI give it a vast breadth of potential application areas, including other areas of scientific and technology research – raising the possibility of an explosion of AI-augmented research rapidly transforming multiple fields of inquiry. Third, many currently prominent AI algorithms and applications, particularly those involving deep learning and reinforcement learning, are opaque in their internal operations, such that it is difficult even for experts to understand how they work.[6] Even distinguished practitioners of these methods have expressed concern that recent advances are unsupported by general principles, often dependent on *ad hoc* adjustments in specific applications, and generative of outputs that are difficult to explain.[7] Two prominent commentators characterized the state of the field as alchemical.[8]

One consequence of these challenges is uncertainty over the causes and implications of recent large advances. Do they represent foundational advances that put general understanding and reproduction of intelligence within reach?[9] Or do

they just reflect the result of continuing advances in several parallel areas – data, algorithms, computational capacity – that are important in practical power, without necessarily representing major scientific breakthroughs or auguring imminent advances in dissimilar problems or reproducing general intelligence?[10] Expert surveys estimating how soon such major landmarks will be achieved show a wide range of views.[11]

A second consequence is deep, high-stakes uncertainty about AI's societal impacts, risks, and governance responses – arguably even more than in other parallel areas of rapid, high-stakes, contentious technologies. This deep uncertainty is driven both by uncertainty about the rate and character of future technological advances in AI, and by uncertainty about how people, enterprises, and societies will interact with these rapidly advancing capabilities: how new capabilities will be used and applied; how people and organizations will react and adjust; how capabilities will further change in response to these adjustments; and how, and how well, societal institutions will manage the consequences to promote the beneficial, limit or mitigate the harmful, and decide in time – competently, prudently, legitimately – which consequences are beneficial and which harmful.

AI Impacts and Governance: Major areas of present inquiry

In the face of this deep uncertainty, current study and speculation on AI impacts and governance has a few salient characteristics. In some respects these are similar to characteristics often found in rapidly growing fields of inquiry and commentary, yet they also reflect the distinct, perhaps unique, characteristics of AI. In broad terms, the approach of many researchers and commentators to AI reflects their prior

interests, concerns, capabilities, and disciplinary traditions. Given the diffuse and labile character of AI and its applications, this is both familiar and sensible. Indeed, it is a phenomenon so familiar as to be nicely captured by old aphorisms and folk tales. AI is the elephant, and we who are trying to reason through its societal impacts and potential responses are the blind men, each feeling around one piece of it.[12] Or alternatively, we all come to the problem of AI with our various forms of hammers, so it looks like a nail.

Attempting to give aggregate characterizations of such a heterogeneous and rapidly growing field is risky, yet necessary. Much present commentary falls into two clusters, mainly distinguished by the immediacy of the concerns they address. The first and larger cluster examines extant or imminently anticipated AI applications that interact with existing legal, political, or social concerns. Prominent examples include how liability regimes must be modified to account for AI-supported decisions (e.g., in health care, employment, education, and finance), or AI-embedded physical objects that interact with humans (autonomous vehicles, robots, internet-of-things devices);[13] racial, gender, or other biases embedded in algorithms, whether in high-stakes decision settings[14] or in more routine but still important matters like search prompts or image labeling;[15] defending privacy under large-scale data integration and analysis;[16] and transparency, explainability, or other procedural values in AI-enabled decisions.[17]

The second cluster of current work concerns the existential risks of extreme AI advances, whether characterized as progressing to general superintelligence, a singularity beyond which AI controls its own further advances, or in other similar terms. Although this perspective is less frequent in scholarly work, it draws support

from the argument that risks of catastrophic or existential consequence merit examination even if they appear temporally distant or unlikely.[18] These prospects loom large in press and popular treatments of AI[19] – to such a degree that researchers arguing for the importance of more immediate risks initially faced an uphill battle gaining similar levels of attention to these concerns.[20]

Between these two clusters lies a wide gap receiving less attention: potential impacts, risks, and governance challenges that are intermediate in time-scale and magnitude, lying between the near-term and the existential. Some scholarship does target this intermediate zone, typically by examining impact mechanisms that are of both immediate and larger-scale, long-term significance. For example, many studies aim to identify technical characteristics of AI systems likely to make them either riskier or more benign.[21] In addition, certain domains and mechanisms of AI social impact – for example, livelihood displacement,[22] social reputation or risk scoring, and autonomous lethal weapons – are of both immediate concern and larger future concern as applications expand. The broad diversity of methods and foci of inquiry in AI impacts is a sensible response to present deep uncertainty about AI capabilities and impacts, and there is great value in this diversity of work. This is a context where interdisciplinarity is at a premium, and the most significant advances are likely to come not just from deep specialization, but also from unexpected connections across diverse fields of inquiry.

The AI PULSE project at UCLA School of Law: A focus on mid-range impacts

In this busy, diverse, and rapidly growing space, the AI PULSE Project at UCLA Law

is a newcomer. Like many others, we aim to advance understanding on the societal impacts of current and anticipated advances in AI, the good, bad, and dangerous; the causes and mechanisms of those impacts; and potential law and other governance responses to inform decision-making, both within traditional legal and regulatory settings, and in new institutional mechanisms and settings.

In looking for areas on which to focus our attention and permit us to make useful contributions, we have used two loose and provisional criteria. First, we have sought issues and questions that can effectively draw on our prior expertise in law and policy, and on prior experiences with other technology law and policy areas such as energy, environment, internet governance, and cybersecurity. We aim to attend carefully to points where scientific or technology matters interact strongly with societal or governance issues – not aiming to focus centrally on technology, which is not our comparative advantage, but rather to recognize and connect with needed expertise via collaborators. Second, we have looked for areas of potential importance that are receiving relatively less attention, and where there is less risk of simply re-treading familiar ground.

This orientation has led us to the intermediate scale of AI impacts, time horizons, and implications, as outlined above. We do not focus principally on immediate concerns already attracting substantial study and attention, nor on existential endpoints. This is not because we judge these uninteresting or unimportant – they are emphatically not – but because so much excellent work is already being done here. This intermediate range, roughly defined in some combination of time-scale and of intensity of impacts and potential disruptions, is unavoidably a bit diffuse in its boundaries. We characterize it by rough conditions that separate it from immediate concerns, and from singularity, existential, or ultimate concerns.

by: Edward Parson, Richard Re, Alicia Solow-Niederman and Elana Zeide

We propose one principal criterion to distinguish this middle range from the applications, consequences, and concerns that characterize the large volume of important work being done on current and near-term AI challenges. AI applications now deployed and in advanced development sit within the context of existing configurations of decision-makers with associated capabilities, interests, and goals. They are being embedded in commercial products and services marketed by existing firms to identified consumers and businesses. They are supporting, and may in some applications replace, human expertise and agency in existing decisions now taken by individual humans, in a wide variety of professional and employment settings – e.g., drivers, machine tool operators, pharmacists, stockbrokers, librarians, doctors, and lawyers. And they are similarly supporting, advising, and perhaps replacing current decisions now made by groups or organizations – i.e., actors larger than one person – but still recognized, abstracted, and sometimes held accountable as an individual, more or less human-like actor, such as corporations, courts, boards, offices, or departments.

But the fact that current AI applications are presently slotting into these existing decisions by existing actors is a social and historical contingency that reflects immediate opportunities to deploy, and sell, AI-infused products and services. There is no reason to expect that AI's capabilities, or its future applications, will necessarily follow the same patterns. The same characteristics that pose challenges to the prediction and governance of AI's societal impacts – its diffuse, labile character, fuzzy boundaries, broad connections to other technologies and fields of inquiry, and foundational nature – also suggest that it is capable of doing things of greater scale, scope, novelty, or complexity than any presently identified decision by a presently identified actor.

It is this greater scale of application, along with associated changes in scope,

complexity, and integration, that we propose as the principal criterion distinguishing near-term impacts and governance challenges from medium-term ones. In this hypothesized mid-range, AI is, at a minimum, doing things that to some degree resemble functions being discharged by present actors, but which due to greatly expanded scale or scope are qualitatively changed in their impacts, for example by divesting current decision-makers of power or fundamentally transforming their aims and inter-relationships. Alternatively, AI might be doing things that are not presently done by any identified single actor, but by larger-scale social processes or networks of multiple actors and institutions – e.g., markets, normative systems, diffuse non-localized institutions, and the international system. Deployed in such settings, AI would take outcomes that are now viewed as emergent properties, equilibria, or other phenomena beyond the reach of any individual decision or centralized control, and make them subject to explication, intentionality, or control. Or as a third alternative, AI might be deployed to carry out actions that are not presently done at all, for various reasons – including that they are beyond the ability of any individual actor to imagine, perceive, or carry out, yet at the same time are not the objects of any linked systems of multi-actor decision-making. In any such settings, we expect the societal impacts and disruptions/transformations of AI, and the associated challenges of governance (indeed, the meaning of governance) to be profoundly transformed, in scale, meaning, and possibly also speed.

Yet this is not the singularity.[23] We also distinguish this middle range from existential or singularity-related risks and by the limitation that AI is not self-directed or independently volitional, but rather is still to a substantial degree developed and deployed under human control. Of course, the practical extent of human control in specific applications may be ambiguous, and the details matter a lot. Moreover, as noted above, even with AI not fully autonomous but practically, or formally, under

human control, there may still be transformative impacts, including vast public as well as private effects and the potential for large-scale disruptions and harms – in addition to the large benefits that are intended and anticipated.

This intermediate range is thick with potential areas and mechanisms of high-stakes AI impacts. In addition to those noted above, involving mechanisms of influence already operating but subject to transformation from increased scale and speed (e.g., livelihood displacement), there are multiple other possibilities. These are substantially more heterogeneous than those already evident in present practice. Even brief reflection suggests a wide range of potential AI applications, impacts, bases for concern, and associated governance challenges. We outline a few below. Many others, similarly plausible, could readily be generated.

- AI as a manager and coordinator of economic functions at larger scale than the scope typical of current enterprises producing and selling goods and services;
- AI as a component of, and to varying degrees supplanting, various forms and functions of state decision-making – legislative, executive, administrative, judicial, or electoral.[24] Small-scale instances of this, and more ambitious proposals, already abound. Implications are profound, both for material outcomes, such as substance of decisions, character and quality of service and program delivery, efficiency and cost; and for legal processes, associated rights, and political principles.
- AI as a disruptor of competitive relations, in commercial competition and other domains of competitive interactions, with the prospect of major shifts in the distribution of power among individual, commercial, public, and other institutions, perhaps also driving large changes in the meaning and determinants of such power.
- AI as an enabler of increased effectiveness and new forms of influence over people in commercial, political, and other societal contexts. Early signs, in 2016 and since, of the potency of algorithmic targeting and tuning of messages for both domestic

and international political campaigns suggest much broader and more potent tools of influence to come. These of course can include AI-empowered influence and manipulation for conventional commercial and political ends, perhaps allowing greater concentration of power than has been possible by other means: one major recent report identified "robust totalitarianism" as one of the major social and political risks of AI.[25] But multiple other forms of influence are possible, including manipulation of behavior, preference, or values, with aims and effects ranging from the benign to the malign. Consider, for example, the possibilities of AI-enabled manipulation for purposes of healthier living, other forms of self-improvement, civic virtue, or conformity, docility, or subordination to the will of another; AI-empowered psychotherapy, new political or social movements, new religions (the first church centered on AI is already established),[26] or cults.

- AI as a scrambler of human individual and collective self-conception, including possibly undermining basic assumptions on which liberal capitalist democratic states and societies are founded – either truly altering the assumed conditions, or altering large-scale confidence or belief in them through increasingly powerful tools of deception such as "deep fakes."[27] These shared assumptions, which include both positive and normative beliefs, exercise a powerful influence on the acceptance and viability of democratic government and market economies. The potential implications of their disruption, for practice of democratic government, and for social cohesion and identity, are profound.

These and similar potential intermediate-range impacts can be read to varying degrees as benign or malign. Many are most likely some of each. But they are clearly large and novel enough to require careful study and assessment, without a strong prior presumption that they predominantly fall one way or the other. Although it may

be tempting to view any large change as harmful, at least initially, we do not presume that the societal impacts of AI will be entirely, or even predominantly, harmful. In fact, there are strong grounds to anticipate large benefits. Innovations in commercial settings generally pursue improved production processes or products and services that people want. We thus expect that the initial, direct effects of AI deployment will mainly be benefits, since they will be realized through voluntary commercial transactions.[28] The axiom of revealed preference is not the last word on societal values, in view of imperfect knowledge and anticipation of individual welfare, identified pathologies of choice, emergent collective outcomes, constrained choices, and manipulation – but neither is it to be arbitrarily rejected. Moreover, even outside purely commercial transaction, a few areas of AI application hold clear prospects for large societal benefits, including medical diagnosis and treatment, scientific and technical research, and environmental monitoring, management, and restoration. Even when technical advances bring some harm – canonically, the harm that incumbents suffer when their business is disrupted by innovations – these local disruptions are often consistent with larger-scale economic and societal benefits.

Yet at the same time, there are well-founded grounds for concern. The greater the new technical capabilities and resultant transformations, the less the resultant disruptions are likely to be confined to the commercial sphere and the more they are likely to implicate larger-scale network and public effects, and political, social, and ethical values separate from those of the market. Moreover, many applications of AI will be in non-market contexts – whether in classic areas of state decision-making or something brand new. In these other contexts, the comforting invisible-hand assumptions about the consonance of individual self-interested choice and aggregate societal benefit do not apply. Individuals and firms developing and applying AI may be sensitive to these broader impacts, but lack the scale of view, knowledge, or

authority to address them effectively. Doing so will require some larger scale apparatus, perhaps including deployment of state regulatory authority, perhaps involving multi-party collaboration among state and non-state actors, including enterprises, technical associations, educational institutions, research funders, scientific and professional associations, civil society organizations, and governmental decision-makers at multiple levels.

Even focusing on this middle range rather than more distant and extreme possibilities, it is difficult to do research about future conditions, risks, or requirements. Future events, conditions, and outcomes are not observable, except via proxies, trends, or enduring conditions and relationships in the present or past. Saying anything useful about future effects unavoidably requires speculative reasoning and acknowledgement of uncertainty, and also disciplining that speculation by reference to current knowledge. Present conditions and trends, scientific knowledge about enduring mechanisms of causation and influence in the world, and the properties and limits of current technologies, all provide useful inputs to reasoning about future conditions and possibilities, but with limits. Structured methods for exploration and contingency planning such as model projections, scenario exercises, and robust decision-making approaches can help stimulate the needed balance of imagination and discipline, but do not surmount deep uncertainties. These challenges are all particularly acute in reasoning through potential impacts, risks, and governance responses for AI, in view of the uniquely challenging characteristics of AI noted above.

One promising way to get insights into AI impacts and responses over this intermediate scale is – instead of focusing on present applications or the technical properties of algorithms and associated data and tools – to focus on decisions: the decisions to develop and refine capabilities, train and test them, and deploy them

for specific purposes in specific settings. Focusing on decisions in turn implies focusing on the actors who make those decisions, whether individuals, teams, or private or public organizations, and the factors that influence their decisions: the interests they seek to advance; the capabilities and resources that characterize their decision choice sets and associated constraints; and the strategic environment in which they make these decisions, including interactions with other decision-makers with resultant emergent properties.

A survey of present AI applications and actors suggests a wide range of interests may be motivating development, adoption, and application decisions. The principal actors developing new methods and capabilities are private firms and academic researchers. But the firms are not typical commercial firms. Several of the leading firms are so big and rich, and so secure in dominant positions in current markets, that they are able to invest in speculative research some distance from commercial application.[29] Even accounting for firms' ostensible profit motives, their need to recruit and retrain scarce, often idiosyncratic, top-rank talent may also shift their mix of interests to some degree, to include scientific interest, technological virtuosity, plus whatever other ambitions or commitments motivate this talent pool.[30] Motivations become even more diverse on the international scale. Several market-leading AI firms are based in China, under varying degrees of government influence – suggesting a blurring of lines between commercial competition and state objectives, both domestic (securing and sustaining political control) and international (geopolitical rivalry, through multiple channels not limited to military). Moreover, not all the major developers are for-profit firms. One major AI enterprise is organized as a not-for-profit, with a stated commitment to development of AI capabilities for social benefit – although the practical implications of this different commitment are

not yet fully evident.[31]

In considering medium-term possibilities for AI applications, risks, and other impacts, it is thus necessary to consider a range of interests and incentives: in addition to commercial competition, potentially important interests might include political competition through electoral or other channels; competition for fame, notoriety, or influence; pursuit of technological advance for its intrinsic pleasures and for professional status; and multiple forms of rivalrous or malicious aims, among individuals, firms, other organizations, and states.[32] This broad list of interests implicates harm mechanisms associated with intentional design choices as well as unforeseen accidents, or design or application failures.

The papers collected here represent the results of an early attempt to examine AI impacts, risks, and governance from this actor and decision-centered perspective. They are based on a May 2018 workshop, "AI in Strategic Context," at which preliminary versions of papers were presented and discussed. In extending their workshop presentations into these papers, we have asked authors to move a little outside their normal disciplinary perspectives, with the aim of making their papers accessible both to academic readers in other fields and disciplines, and to sophisticated policy-oriented and other non-academic audiences. In the spirit of the larger-scale project, we also encouraged authors to be a little more speculative than they would normally be able to do when writing for their scholarly peers.

This collection includes seven of the resultant papers, spanning a broad range of applications, impacts, disciplinary perspectives, and views of governance. Sterbenz and Trager use an analytic approach rooted in game theory to characterize the effects of a particular class of AI, autonomous weapons, on conflicts and crisis-escalation situations. Ram identifies a new potential mechanism of harmful impact

based on increasing pursuit of the technical capability, "one-shot learning." She argues that increased use of one-shot learning might exacerbate present concerns about both bias and opacity, and assigns a central role to trade secrecy in these harms. Grotto distills a set of concrete lessons for governance of AI by analogizing to a prior political conflict over another high-stakes and contentious technology, GMOs, drawing particularly on the variation provided by disparate policy outcomes in the United States and the European Union. Marchant argues both that effective governance of AI is needed and that conventional governmental regulation is unlikely in the near term, and instead proposes a soft-law approach to governance, including a specific institutional recommendation to address some of the most obvious limitations of soft-law governance. Moving from regulatory practice to political theory, Panagia explores the alternative ways algorithms can be regarded as political artifacts. He argues that algorithms are devices that order the world, in relations between people and objects and among people, and that this is a distinct, and broader, role than that conventionally proposed by critical thinkers of the left, who tend to view algorithms predominantly as instruments of political domination among people. Osoba offers a provocative sketch of a new conceptual framework to consider global interactions over AI and its governance, arguing that there exist distinct linked technical and cultural systems, or "technocultures," around AI, and that current diversity of technocultures is likely to be a persistent source of regulatory friction among regional governance regimes. And finally, Lempert provides a provocative vision of the capability of AI to drive large-scale divergence in human historical trajectories. Drawing on historical analogies explored in the scholarship of Elizabeth Anderson, he presents two scenarios in which large-scale AI deployment serves either to restrict or to strengthen human agency, and speculates on the technical structures of algorithms that might tend to nudge a world deploying them toward those two divergent futures.

by: Edward Parson, Richard Re, Alicia Solow-Niederman and Elana Zeide

Closing Observations:

These papers represent the first output from what we aim to make a continuing inquiry. In our view, they nicely illustrate what needs to happen to build a coherent and progressive body of work in a new and diffuse, but high-stakes area: targeting interesting and important questions and effectively deploying discipline-based concepts and frameworks, yet also striving for clear communication across disciplinary lines and accessibility both outside the field and outside the academy – while still maintaining precision. At the same time, they also illustrate the challenges facing this area of inquiry: its vast breadth, the diversity of relevant disciplinary languages and perspectives – and the difficulty of directing focus forward beyond immediate concerns, of being willing to engage in speculation yet keeping the speculation disciplined and connected to current knowledge and expertise. These are the aims and the challenges of the project, which we plan to further explore in additional workshops, collaborations, and publications.

By Edward Parson, Dan and Rae Emmett Professor of Environmental Law, Faculty Co-Director of the Emmett Institute on Climate Change and the Environment, and Co-Director of the AI Law and Policy Program, UCLA School of Law, UCLA Dept

Richard Re, Assistant Professor Co-Director of PULSE and the AI Law and Policy Program

Alicia Solow-Niederman, PULSE Fellow in AI Law and Policy, UCLA School of

by: Edward Parson, Richard Re, Alicia Solow-Niederman and Elana Zeide

Law

and Elana Zeide, PULSE Fellow in AI Law and Policy, UCLA School of Law.

1. *See* surveys reported in Vincent C. Müller & Nick Bostrom, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, https://nickbostrom.com/papers/survey.pdf; a more recent survey at Katja Grace, et al., *When Will AI Exceed Human Performance?: Evidence from AI Experts*, https://arxiv.org/pdf/1705.08807.pdf. See also Janna Anderson, Lee Rainie & Alex Luchsinger, *Artificial Intelligence and the Future of Humans*, Pew Research Center (Dec. 10, 2018), http://www.pewinternet.org/2018/12/10/artificial-intelligence-and-the-future-of-humans/ and a survey by Martin Ford discussed in James Vincent, *This Is When AI's Top Researchers Think Artificial General Intelligence Will Be Achieved*, The Verge (Nov. 27, 2018, 1:05 PM), https://www.theverge.com/2018/11/27/18114362/ai-artificial-general-intelligence-when-achieved-martin-ford-book.
2. *See, e.g.,* Harvey Brooks, *Technology Assessment in Retrospect*, 17 Science, Technology & Human Values 17-29 (1976); David Collingridge, Social Control of Technology (1978); Langdon Winner, *Do Artifacts Have Politics?*, 109 Daedalus 1 (1980); Sheila Jasanoff, Gerald E. Markle, James C. Peterson, & Trevor J. Pinch, Handbook of Science and Technology Studies (1995); Frank Geels, Technological Transitions and System Innovations: A Co-evolutionary and Socio-technical Analysis (2005).
3. *See* Winner, *supra* note 2.

4. *See* Collingridge, *supra* note 2.
5. *See* Edward A. Parson, *Social Control of Technological Risks: The Dilemma of Knowledge and Control in Practice, and Ways to Surmount It*, 64 UCLA L. Rev. Disc. 464 (2016).
6. *See, e.g.*, Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 1085 (2018); Will Knight, *The Dark Secret at the Heart of AI*, MIT Technology Review (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai; *See also* David Gunning, Explainable Artificial Intelligence, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence. *But see* Chris Olah, et al., *The Building Blocks of Interpretability*, Distill (Mar. 6, 2018), https://distill.pub/2018/building-blocks/; David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC Davis L. Rev. 653 (2017).
7. *See* Matthew Hutson, *AI Faces a Reproducibility Crisis*, 359 Science 725-726 (2018); Paul Voosen, *The AI Detectives*, 357 Science 22-27 (2017); D. Sculley, Jasper Snoek, Ali Rahimi,& Alex Wiltschko, *Winner's Curse: On Pace, Progress and Empirical Rigor* Open Review (2018), https://openreview.net/pdf?id=rJWF0Fywf.
8. Ali Rahimi & Ben Recht, *Reflections on Random Kitchen Sinks*, Arg Min Blog (Dec. 5, 2017), http://www.argmin.net/2017/12/05/kitchen-sinks/ (transcript of talk delivered at NIPS 2017), further elaborated in Ali Rahimi & Ben Recht, *An Addendum to Alchemy*, Arg Min Blog (Dec. 11, 2017), http://www.argmin.net/2017/12/11/alchemy-addendum/. But see Yann LeCun (@yann.lecun), Facebook (Dec. 6, 2017, 8:57 AM), https://www.facebook.com/yann.lecun/posts/10154938130592143. The characterization of AI as resembling alchemy has an extensive history, having first been articulated decades before the recent flood of ML-driven advances. *See, e.g.*,

Hubert L. Dreyfus, *Alchemy and Artificial Intelligence*, RAND Corporation Paper P-3244, (1965), https://www.rand.org/pubs/papers/P3244.html

9. *E.g.*, Pedro Domingos, The Master Algorithm (2015).
10. *See* Adnan Darwiche, *Human-Level Intelligence or Animal-Like Abilities?*, arXiv (Jul. 2017), https://arxiv.org/abs/1707.04327. *See also* Judea Pearl & Dana MacKenzie, The Book of Why (2018).
11. *Cf. supra* note 1.
12. *Tittha Sutta (Ud.6.4)*, in The Middle Length Discourses of the Buddha. (B. Nanamoli and B. Bodhi tr., Buddhist Texts Society 2015). *See also* The Blind Men and The Elephant: collected poems of John Godfrey Saxe, James R. Osgood and Company (1876).
13. *See, e.g.*, George S. Cole, Tort Liability for Artificial Intelligence and Expert Systems, 10 Computer L.J. 127 (1990); Nathan A. Greenblatt, *Self-driving Cars and the Law*, IEEE Spectrum (Feb. 2016), at 42; John Kingston, *Artificial Intelligence and Legal Liability*, arXiv (Feb. 21, 2018), arxiv.org/abs/1802.07782; Andrew Selbst, *Tort Law's AI Problem* (draft ms., forthcoming).
14. *See, e.g.*, Julia Angwin, et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. *See also* Ed Yong, *A Popular Algorithm is No Better at Predicting Crimes Than Random People*, Atlantic (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/. *But see* William Dieterich, et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, Northpointe Inc. Research Department (July 8, 2016), https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final

-070616.html and Sam Corbett-Davies, et al. *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*. Washington Post (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.d4b73eb99c06; https://arxiv.org/pdf/1811.00731.pdf. *ProPublica has responded to Northpoint's critique. Julia Angwin & Jeff Larson, ProPublica Responds to Company's Critique of Machine Bias Story*, ProPublica (July 29, 2016), https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story. *See also* Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016); Solon Barocas, Moritz Hardt, and Arvind Naranayan, Fairness in Machine Learning (2018), http://fairmlbook.org; Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, & Aaron Roth *Fairness in Criminal Justice Risk Assessments: The State of the Art*, arXiv (Mar. 27, 2017), https://arxiv.org/abs/1703.09207.

15. *See, e.g.*, Safiya Umoja Noble, Algorithms of Oppression (2018); Algorithmic Justice League, https://www.ajlunited.org. *See also* James Vincent, *Google 'Fixed' It's Racist Algorithm by Removing Gorillas from its Image-Labeling Tech*, The Verge (Jan. 12, 2018, 10:35 AM), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.
16. *See, e.g.*, Omar Tene & Jules Polonetsky, *Privacy in the Age of Big Data*, Stan. L. Rev. (Feb 2012), https://www.stanfordlawreview.org/online/privacy-paradox-privacy-and-big-data/. *See also* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010); Arvind Narayanan & Vitaly

Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 Proc. of IEEE Symp. on Security & Privacy 111; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671 (2016).

17. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 1085 (2018); Ryan Budish, et al., *Accountability of AI Under the Law: The Role of Explanation*, Berkman Klein Center, https://arxiv.org/pdf/1711.01134.pdf.
18. *See, e.g.*, Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (2014); Ray Kurzweil, The Singularity is Near: When Humans Transcend Biology (2005); Nick Bostrom, *Existential Risks*, 9 Journal of Evolution and Technology 1, 1-31 (2002); Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, *in* Global Catastrophic Risks 308-45 (Nick Bostrom & Milan irkovi , eds., 2008). *See also* Rohin Shah, *Alignment Newsletter*, https://rohinshah.com/alignment-newsletter.
19. *See, e.g.*, James Barrat, Our Final Invention: Artificial Intelligence and the End of the Human Era(2013). *See also* Michael Shermer, *Apocalypse AI*, Scientific American, Mar. 2017, at 77; Raffi Khatchadourian, *The Doomsday Invention: Will Artificial Intelligence Bring Us Utopia or Destruction?* The New Yorker, 23 Nov. 2015; Henry Kissinger, *How the Enlightenment Ends*, Atlantic Monthly, June 2018.
20. *See, e.g.*, Kate Crawford & Ryan Calo, *There Is a Blind Spot in AI Research*, 538 Nature 311-313 (2016).
21. *See, e.g.*, Eliezer Yudkowsky, *Complex Value Systems in Friendly AI* in International Conference on Artificial General Intelligence (Aug. 2011) at 388-393. Future of Life Institute, *An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence*, https://futureoflife.org/ai-open-letter; Eliezer Yudkowsky (2001), Creating Friendly AI 1.0. at https://intelligence.org/files/CFAI.pdf; Mark Waser, *Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and*

Morals for Intelligent Machines (Including Humans), 71 Procedia Computer Science 106–111 (2015); Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, & Dan Mané, *Concrete Problems in AI Safety*, arXiv(July 25, 2016), https://arxiv.org/abs/1606.06565.

22. *See, e.g.*, NBER Economics of Artificial Intelligence Conference 2018, https://www.economicsofai.com/nber-conference-2018/; Economics of AI, https://www.economicsofai.com/nber-conference-toronto-2017/. *See also* Carl B. Frey and Michael A. Osborne, *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, Oxford Martin Programme on the Impacts of Future Technology (Sept. 17, 2013), https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.
23. *Cf. supra* note 2.
24. *See, e.g.*, Eugene Volokh, *Chief Justice Robots*, Duke L.J. (forthcoming 2019), https://reason.com/assets/db/15474336323990.pdf.
25. Allan Dafoe, *AI Governance: A Research Agenda*, Future of Humanity Institute (Aug 27, 2018), https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf.
26. *See* Way of the Future Church, *What Is This All About?*, http://www.wayofthefuture.church/. *See also* Paul Golata, *The Church of Artificial Intelligence*, Land Center (Feb. 4, 2018), https://thelandcenter.org/the-church-of-artificial-intelligence-the-way-of-the-future-denies-the-way-the-truth-and-the-life/.
27. *See* Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Crisis for National Security, Democracy, and Privacy?*, Lawfare (Feb. 4, 2019), https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy.
28. Though these benefits may accrue unevenly or in fact only be a boon to those who

can afford to engage in these commercial transactions.

29. See Nathan Benaich & Ian Hogarth, *State of AI* (June 29, 2018), https://www.stateof.ai/ (summarizing AI intellectual property concentration among eight global companies: Alibaba, Amazon, Apple, Baidu, Facebook, Google, Microsoft, and Tencent).
30. Consider accounts of walkouts at multiple technology firms protesting proposals to bid on a military contract. E.g., Alexia Fernández Campbell, "How Tech Employees Are Pushing Silicon Valley to Put Ethics Before Profit," Vox (Oct 18, 2018), https://www.vox.com/technology/2018/10/18/17989482/google-amazon-employee-ethics-contracts.
31. See About OpenAI, OpenAI, https://openai.com/about/.
32. *See* Miles Brundage et al., *The Malicious Use of AI: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute (Feb. 2018), https://maliciousaireport.com/.