



AGTech Forum Briefing Book: State Attorneys General and Artificial Intelligence

Citation

Hessekiel, Kira, Eliot Kim, James Tierney, Jonathan Yang, and Christopher T. Bavitz. 2018. AGTech Forum Briefing Book: State Attorneys General and Artificial Intelligence, May 8-9, 2018, Harvard Law School. Berkman Klein Center for Internet & Society.

Published version

<https://cyber.harvard.edu/publications/2018/05/AGTech>

Link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:37184705>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)



AGTECH

FORUM

BRIEFING
BOOK

State Attorneys General
and Artificial Intelligence

MAY 8 & 9 2018

HARVARD LAW SCHOOL

Kira Hessekiel, Eliot Kim,
James Tierney, Jonathan Yang,
& Christopher T. Bavitz

Part of the Ethics and
Governance of AI Initiative





Introduction

Artificial intelligence is already starting to change our lives. Over the coming decades, these new technologies will shape many of our daily interactions and drive dramatic economic growth. As AI becomes a core element of our society and economy, its impact will be felt across many of the traditional spheres of AG jurisdiction. Members of AG offices will need an understanding of the AI tools and applications they will increasingly encounter in consumer devices, state-procured systems, the court system, criminal forensics, and others areas that touch on traditional AG issues like consumer privacy, criminal justice, and representing state governments.

The modest goal of this primer is to help state AGs orient their thinking by providing both a broad overview of the impact of AI on AG portfolios, and a selection of resources for further learning regarding specific topics. As with any next technology, it is impossible to predict exactly where AI will have its most significant on matters of AG jurisdiction. Yet AGs can better prepare themselves for this future by maintaining a broad understanding of how AI works, how it can be used, and how it can impact our economy and society. In success, AGs can play a key constructive role in preventing misconduct, shaping guidelines, and ultimately maximizing the positive impact of these exciting new technologies. We intend for this briefing book to serve as a jumping-off point in that preparation, setting a baseline of understanding for the AGTech Forum and providing resources for specific learning beyond our workshop.

What is Artificial Intelligence (AI)?

AI is an umbrella term for a set of technologies that replicate human behavior, allowing computers to mimic, supplement, and in some cases replace human direction. Though there is no one universal definition, a 2016 White House report highlights the following taxonomy from a popular AI textbook:

1. systems that think like humans (e.g., cognitive architectures and neural networks)
2. systems that act like humans (e.g., pass the Turing test via natural language processing; knowledge representation, automated reasoning, and learning)
3. systems that think rationally (e.g., logic solvers, inference, and optimization)
4. systems that act rationally (e.g., intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting).¹

¹ Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd Edition) (Essex, England: Pearson, 2009).

Some common AI techniques include machine learning, deep learning, and neural networks, each of which try to draw connections between data points—ranging from demographic data to live visual feeds to email archives—to arrive at insights, often incorporating more information than human analysis could digest. Many AI tools look for patterns in past behavior and phenomena, and are thus shaped as much by the data on which they are “trained” as they are by their programming.

While the term AI may bring to mind images of killer robots, most of today’s AI applications will

primarily allow companies and individuals to do many of the same things they currently do (e.g. make decisions, detect patterns, identify potential clients) better, faster, and on a wider scale. These tools are commonly classified as “Narrow AI,” and are narrowly tailored to the performance of specific tasks, such as assembly line management, advanced ad targeting, medical and academic research, image recognition, and many other familiar activities. In contrast, “General AI” theoretically would completely replicate or even exceed human intelligence. Context is especially key with respect to AI; the same algorithm can be more or less problematic depending on the quality of the data fed into it, the degree of reliance the company places on its conclusion, and the type of decisions to which it is applied.

Starting around 2010, breakthroughs and increased investment in several new computational techniques led to its rapid incorporation in numerous commercial applications, a trend that continues in earnest today. A 2016 White House report points to three cumulative drivers of progress in this space: 1) the increased availability of “big data” from many different sources, which 2) provided raw material for dramatically improved machine learning approaches and algorithms, which 3) relied on the capabilities of much more powerful computers. This acceleration is likely to continue, especially as AI gradually advances beyond the narrowly tailored tasks its current iterations can currently perform.

How might it affect the work of State AGs?

AI will increasingly impact many of the actors and activities that state AGs already represent and regulate. Any organization can potentially use AI to support its work; even smaller and less sophisticated actors will increasingly be able to utilize some form of AI in their activities. Consumer-facing companies will be able to target ads, adjust prices and offerings, and otherwise tailor their interactions with customers better than ever before. Autonomous vehicles and other hardware may put physical infrastructure and residents at increased risk of accidents and systemic failure. AI will enable even more accurate simulations of human interaction, creating the potential both for better customer/constituent relations, and for more sophisticated fraud. At the structural level, companies may be able to leverage early advantages in AI technology—or access to large amounts of data—to entrench their market dominance and drive out competition. These applications will create a hunger for access to more data from more people, putting constant pressure on privacy regulators.

As AI spreads into traditional spheres of concern, AGs will in turn need to understand what these new technologies can—and cannot—do. To effectively regulate AI without unduly constraining its many beneficial applications, for both private and public sector alike, AGs should be neither complacent nor paranoid about the impacts AI will have. The broader impacts of AI on the economy and society will also generate greater pressure on AGs to

investigate and regulate actors who use AI, requiring AGs to exercise their discretion to distinguish actionable complaints from generalized objections to new technology.

AGs must also rethink how they investigate, litigate, and settle enforcement matters, as the spread of AI shifts decisions from people to proprietary algorithms, requiring sufficient familiarity with AI technologies. At the tactical level, an understanding of AI will help AGs prioritize their efforts. For example, a civil investigative demand for a company's proprietary algorithm will almost certainly be fiercely resisted by the target, and may be neither necessary or even helpful to identifying potential wrongdoing. Instead, an AG may get more out of learning how the data set used by the algorithm was assembled. AI may allow bad actors to coordinate in indirect ways that are harder to detect, let alone prove. These distinctions require an understanding of what role AI played (and did not play) in the behavior under investigation.

AGs will also need to understand how AI can be part of a potential solution to misconduct. The same capabilities that AI brings to bear in targeting ads or identifying patterns could, for example, help a bank identify suspicious internal practices. During Facebook CEO Mark Zuckerberg's 2018 testimony before Congress regarding foreign exploitation of the social network to influence the 2016 election, Zuckerberg cited AI as a potential

tool to help police the content on the platform.² Some misconduct, even if not originally caused by the use of AI, may be best addressed through the careful application of AI tools that can monitor behavior on a vast scale. The better AGs understand these tools, the better they will be able to evaluate—and question—their effectiveness as a part of a settlement.

AGs will also be called upon to support government users of AI. AGs will need to be able to advise state agencies and local governments on the risks and proper use of algorithmic AI tools, ranging from automated constituent-facing chatbots to risk assessment algorithms already in use in criminal sentencing. These government decisions will sometimes be controversial and be subject to challenge. Do sentencing algorithms violate due process or equal protection? Are the resource allocations made by an AI program used by a government agency arbitrary and capricious? As in many other areas, AGs will need to be prepared to distinguish—including in a judicial forum—the AI tools used by government from those used by a private actor the AG is prosecuting down the hall.

² Tim Johnson, “Facebook embraces A.I., and risks further spooking consumers.” McClatchy DC. (April 16, 2018). <http://www.mcclatchydc.com/news/nation-world/national/article208852029.html>

Keeping Up With The Future

AI will transform our world in the coming decades, inevitably arising across the range of issues in the traditional AG portfolio. Many significant applications are only just starting to emerge in the marketplace and in society. As in any area of such opportunity, the enthusiasm of actors and rapid pace of new developments in this space will generate many opportunities for mistakes and misconduct. The more AGs keep up to date with the changing nuances of AI, including its limitations, the more constructive and effective they will be.

AGs can play a pivotal role in shaping how AI is integrated into commerce and society. As in other new and constantly changing fields, such as privacy, industry may look to AGs to help shape guidelines and statutory frameworks. AGs may therefore play a welcome role in this evolving conversation, articulating what kinds of rules and explanations might give the public confidence in the fairness of AI decisionmaking and shaping the adoption of industry standards or public regulations.

As AI is adopted by more and more actors and becomes a growing driver of economic growth, AGs will find themselves on the frontlines. While there will be distinct new challenges posed by AI, many of the potential harms and areas of constructive engagement will fall within familiar areas of AG jurisdiction—from consumer protection to criminal justice to agency representation.

“Technology is not destiny; economic incentives and public policy can play a significant role in shaping the direction and effects of technological change. Given appropriate attention and the right policy and institutional responses, advanced automation can be compatible with productivity, high levels of employment, and more broadly shared prosperity.³

³ Artificial Intelligence, Automation, and the Economy. Executive Office of the President. (December 2016), 3.

Before the AGTech Forum: Learning the Fundamentals

AI and the Law: Setting the Stage

Urs Gasser
Executive Director at the Berkman Klein Center

While there is reasonable hope that superhuman killer robots won't catch us anytime soon, narrower types of AI-based technologies have started changing our daily lives: AI applications are rolled out at an accelerated pace in schools, homes, and hospitals, with digital leaders such as high tech, telecom, and financial services among the early adopters. AI promises enormous benefits for the social good and can improve human well-being, safety, and productivity, as anecdotal evidence suggests. But it also poses significant risks for workers, developers, firms, and governments alike, and we as a society are only beginning to understand the ethical, legal, and regulatory challenges associated with AI, as well as develop appropriate governance models and responses.

Having the privilege to contribute to some of



The Revolution by Fonytas, licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

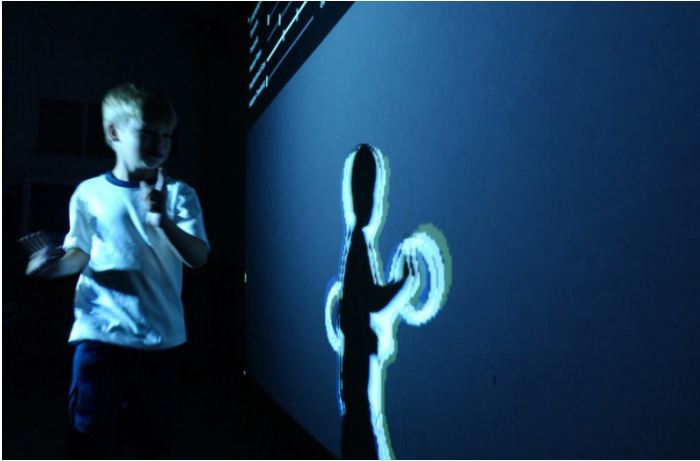
the conversations and initiatives in this thematic context, I plan to share a series of observations, reflections, and points of view over the course of the summer with a focus on the governance of AI. In this opening post, I share some initial thoughts regarding the **role of law** in the age of AI. Guiding themes and questions I hope to explore, here and over time, include the following: What can we expect from the legal system as we deal with both the risks and benefits of AI-based applications? How can (and should) the law approach the multi-faceted AI phenomenon? How can we prioritize among the many emerging legal and regulatory issues, and what tools are available in the toolbox of lawmakers and regulators? How might the law deal with the (potentially distributed) nature of AI applications? More fundamentally, what is the relevance of a law vis-à-vis a powerful technology such as AI? What can we learn from past cycles of technological innovation as we approach these questions? How does law interact with other forms of governance? How important is the role of law in a time where AI starts to embrace the law itself? How can we build a learning legal system and measure progress over time?

I hope this Medium series serves as a starting point for a lively debate across disciplines, boundaries, and geographies. To be sure, what I am going to share in these articles is very much in *beta* and subject to revision and new insight, and I'm looking forward to hearing and learning from all of you. Let's begin with some initial observations.

Find the original version of this article on *Medium*
<https://medium.com/berkman-klein-center/ai-and-the-law-setting-the-stage-48516fdalb11>

Lawmakers and regulators need to look at AI not as a homogenous technology, but a set of techniques and methods that will be deployed in specific and increasingly diversified applications. There is currently no generally agreed-upon definition of AI. What is important to understand from a technical perspective is that AI is not a single, homogenous technology, but a rich set of subdisciplines, methods, and tools that bring together areas such as speech recognition, computer vision, machine translation, reasoning, attention and memory, robotics and control, etc. These techniques are used in a broad range of applications, spanning areas as diverse as health diagnostics, educational tutoring, autonomous driving, or sentencing in the criminal justice context, to name just a few areas of great societal importance. From a legal and regulatory perspective, the term AI is often used to describe a quality that cuts across some of these applications: the degree of autonomy of such systems that impact human behavior and evolve dynamically in ways that are at times even surprising to their developers. Either way, whether using a more technical or phenomenological definition, the justification and timing of any legal or regulatory intervention as well as the selection of governance instruments will require a careful contextual analysis in order to be technically workable and avoid both over-generalization as well as unintended consequences.

Given the breadth and scope of application, AI-based technologies are expected to trigger a myriad of legal and regulatory issues not only at the intersections of data and algorithms, but also of infrastructures and humans. As a growing number of increasingly impactful AI technologies make their way out of research labs and turn into industry applications, legal and regulatory systems will be confronted with a multitude of issues of different levels of complexity that need to be addressed. Both lawmakers and regulators as well as other actors will be affected by the pressure that AI-based applications place on the legal system (here as a response system), including courts, law enforcement, and lawyers, which highlights the importance of knowledge transfer and education (more on this point below). Given the (relative) speed of development, scale, and potential impact of AI development and deployment, lawmakers and regulators will have to prioritize among the issues to be addressed in order to ensure the quality of legal processes and outcomes—and to avoid unintended consequences of interventions. Trending issues that seem to have a relatively high priority include questions around bias and discrimination of AI-based applications, security vulnerabilities, privacy implications of such highly interconnected systems, conceptions of ownership and intellectual property rights over AI creative works, and issues related to liability of AI systems, with intermediary liability perhaps at the forefront. While an analytical framework to categorize these legal questions is currently missing, one might consider a layered model such as a version of the interop “cake model” developed elsewhere in order to map and cluster these emerging issues.



Gesture Recognition by Comixboy, licensed under the Creative Commons Attribution 2.5 Generic license.

When considering (or anticipating) possible responses by the law vis-à-vis AI innovation, it might be helpful to differentiate between application-specific and cross-cutting legal and regulatory issues. As noted, AI-based technologies will affect almost all areas of society. From a legal and regulatory perspective, it is important to understand that new applications and systems driven by AI will not evolve and be deployed in a vacuum. In fact, many areas where AI is expected to have the biggest impact are already heavily regulated industries—consider the transportation, health, and finance sectors. Many of the emerging legal issues around specific AI applications will need to be explored in these “sectoral” contexts. In these areas, the legal system is likely to follow traditional response patterns when dealing with technological innovation, with a default on the application of existing norms to the new phenomenon and, where necessary, gradual reform of existing laws. Take the recently approved German regulation of self-driving cars as an example, which came in the form of an amendment to the existing Road Traffic Act. In parallel, a set of cross-cutting issues is emerging, which will likely be more challenging to deal with and might require more substantive innovation within the legal system itself. Consider for instance questions about appropriate levels of interoperability in the AI ecosystem at the technical, data, and platform

layers as well as among many different players, issues related to diversity and inclusion, and evolving notions of the transparency, accountability, explainability, and fairness of AI systems.

Information asymmetries and high degrees of uncertainty pose particular difficulty to the design of appropriate legal and regulatory responses to AI innovations—and require learning systems. AI-based applications—which are typically perceived as “black boxes”—affect a significant number of people, yet there are nonetheless relatively few people who develop and understand AI-based technologies. This information asymmetry also exists between the technical AI experts on the one hand, and actors in the legal and regulatory systems on the other hand, who are both involved in the design of appropriate legal and regulatory regimes, which points to a significant educational and translational challenge. Further, even technical experts may disagree on certain issues the law will need to address—for instance, to what extent a given AI system can or should be explained with respect to individual decisions made by such systems. These conditions of uncertainty in terms of available knowledge about AI technology are amplified by normative uncertainties: people and societies will need time to build consensus among values, ethics, and social norm baselines that can guide future legislation and regulation, the latter two of which also have to manage value trade-offs. Together, lawmakers and regulators have to deal with a tech environment characterized by uncertainty and complexity, paired with business dynamics that seem to reward time-to-market at all cost, highlighting the importance of creating highly adaptive and responsive legal systems that can be adjusted

as new insights become available. This is not a trivial institutional challenge for the legal system and will likely require new instruments for learning and feedback-loops, beyond traditional sunset clauses and periodic reviews. Approaches such as regulation 2.0, which relies on dynamic, real-time, and data-driven accountability models, might provide interesting starting points.

The responses to a variety of legal and regulatory issues across different areas of distributed applications will likely result in a complex set of sector-specific norms, which are likely to vary across jurisdictions. Different legal and regulatory regimes aimed at governing the same phenomenon are of course not new and are closely linked to the idea of jurisdiction. In fact, the competition among jurisdictions and their respective regimes is often said to have positive effects by serving as a source of learning and potentially a force for a “race to the top.” However, discrepancies among legal regimes can also create barriers when harnessing the full benefits of the new technology. Examples include not only differences in law across nation states or federal and/or state jurisdictions, but also normative differences among different sectors. Consider, for example, the different approaches to privacy and data protection in the US vs. Europe and the implications for data transfers, an autonomous vehicle crossing state boundaries, or barriers to sharing data for public health research across sectors due to diverging privacy standards. These differences might affect the application as well as the development of AI tech itself. For instance, it is argued that the relatively lax privacy standards in China have contributed to its role as a leader in facial recognition technology. In the age of AI,

the creation of appropriate levels of legal interoperability—the working together of legal norms across different bodies and hierarchy of norms and among jurisdictions—is likely to become a key topic when designing next-generation laws and regulations.

Law and regulation may constrain behavior yet also act as enablers and levelers—and are powerful tools as we aim for the development of AI for social good. In debates about the relationship between digital technology and the law, the legal system and regulation are often characterized as an impediment to innovation, as a body of norms that tells people what not to do. Such a characterization of law is inadequate and unhelpful, as some of my previous research argues. In fact, law serves several different functions, among them the role of an enabler and a leveler. The emerging debate about the “regulation of AI” will benefit from a more nuanced understanding of the functions of law and its interplay with innovation. Not only has the law already played an enabling role in the development of a growing AI ecosystem—consider the role of IP (such as patents and trade secrets) and contract law when looking at the business models of the big AI companies, or the importance of immigration law when considering the quest for talent—but law will also set the ground for the market entry of many AI-based applications, including autonomous vehicles, the use of AI-based technology in schools, the health sector, smart cities, and the like. Similarly, law’s performance in the AI context is not only about managing its risk, but is also about principled ways to unleash its full benefits, particularly for the social good—which might require managing adequate levels of open-

ness of the AI ecosystem over time. In order to serve these functions, law needs to overcome its negative reputation in large parts of the tech community, and legal scholars and practitioners play an important educational and translational role in this respect.

Law is one important approach to the governance of AI-based technologies. But lawmakers and regulators have to consider the full potential of available instruments in the governance toolbox. Over the past two decades of debate about the regulation of distributed technologies with global impact, rough consensus has emerged in the scholarly community that a governance approach is often the most promising conceptual starting point when looking for appropriate “rules of the game” for a new technology, spanning a diverse set of norms, control mechanisms, and distributed actors that characterize the post-regulatory state. At a fundamental level, a governance approach to AI-based technologies embraces and activates a variety of modes of regulation, including technology, social norms, markets and law, and combines these instruments with a blended governance framework. (The idea of combining different forms of regulation beyond law is not new and, as applied to the information environment, is deeply anchored in the Chicago-school and was popularized by Lawrence Lessig.) From this ‘blended governance’ perspective, the main challenge is to identify and activate the most efficient, effective, and legitimate modalities for any given issue, and to successfully orchestrate the interplay among them. A series of advanced regulatory models that have been developed over the past decades (such as the active matrix theory, polycentric governance, hybrid regulation,

and mesh regulation, among others) can provide conceptual guidance on how such blended approaches might be designed and applied across multiple layers of governance. From a process perspective, AI governance will require distributed multi-stakeholder involvement, typically bringing together civil society, government, the private sector, and the technical and academic community—collaborating across the different phases of a governance lifecycle. Again, lessons regarding the promise and limitations of multi-stakeholder approaches can be drawn from other areas, including Internet governance, nanotechnology regulation, or gene drive governance, to name just a few.



Innovation by Boegh, Creative Commons Attribution 2.0 Generic license.

In a world of advanced AI technologies and new governance approaches towards them, the law, the rule of law, and human rights remain critical bodies of norms. The previous paragraph introduced a broader governance perspective when it comes to the “regulation” (broadly defined) of issues associated with AI-based applications. It characterized the law as only one, albeit important, instrument among others. Critics argue that in such a “regulatory paradigm,”

law is typically reduced to a neutral instrument for social engineering in view of certain policy goals and can be replaced or mixed with other tools depending on its effectiveness and efficiency. A relational conception of law, however, sees it neither as instrumentalist nor autonomous. Rather, such a conception highlights the normativity of law as an institutional order that guides individuals, corporations, governments, and other actors in society, ultimately aiming (according to one prominent school of thought) for justice, legal certainty, and purposiveness. Such a normative conception of law (or at least a version of it), which takes seriously the autonomy of the individual human actor, seems particularly relevant and valuable as a perspective in the age of AI, where technology starts to make decisions that were previously left to the individual human driver, news reader, voter, judge, etc. A relational conception of law also sees the interaction of law and technology as co-constitutive, both in terms of design and usage—opening the door for a more productive and forward-looking conversation about the governance of AI systems. As one starting point for such a dialogue, consider the notion of society-in-the-loop. Recent initiatives such as the IEEE Global Initiative on Ethically Aligned Design further illustrate how fundamental norms embedded in law might guide the creation and design of AI in the future, and how human rights might serve a source of AI ethics when aiming for the social good, at least in the Western hemisphere.

As AI applies to the legal system itself, however, the rule of law might have to be re-imagined and the law re-coded in the longer run. The rise of AI leads not only to questions about

the ways in which the legal system can or should regulate it in its various manifestations, but also the application of AI-based technologies to law itself. Examples of this include the use of AI that supports the (human) application of law, for instance to improve governmental efficiency and effectiveness when it comes to the allocation of resources, or to aid auditing and law enforcement functions. More than simply offering support, emerging AI systems may also increasingly guide decisions regarding the application of law. “Adjudication by algorithms” is likely to play a role in areas where risk-based forecasts are central to the application of law. Finally, the future relationship between AI and the law is likely to become even more deeply intertwined, as demonstrated by the idea of embedding legal norms (and even human rights, see above) into AI systems by design. Implementations of such approaches might take different forms, including “hardwiring” autonomous systems in such ways that they obey the law, or by creating AI oversight programs (“AI guardians”) to watch over operational ones. Finally, AI-based technologies are likely to be involved in the future creation of law, for instance through “rule-making by robots,” where machine learning meets agent-based modeling, or the vision of an AI-based “legal singularity.” At least some of these scenarios might eventually require novel approaches and a reimagining of the role of law in its many formal and procedural aspects in order to translate them into the world of AI, and as such, some of today’s laws will need to be re-coded.

Thanks to the Special Projects Berkman Klein Center summer interns for research assistance and support.

Accountability of AI Under the Law: The Role of Explanation

Finale Doshi-Velez*, Mason Kortz*,

for the Berkman Klein Center Working Group on Explanation and the Law:

Ryan Budish, Berkman Klein Center for Internet and Society at Harvard University

Chris Bavitz, Harvard Law School; Berkman Klein Center for Internet and Society at Harvard University

Finale Doshi-Velez, John A. Paulson School of Engineering and Applied Sciences, Harvard University

Sam Gershman, Department of Psychology and Center for Brain Science, Harvard University

Mason Kortz, Harvard Law School Cyberlaw Clinic

David O'Brien, Berkman Klein Center for Internet and Society at Harvard University

Stuart Shieber, John A. Paulson School of Engineering and Applied Sciences, Harvard University

James Waldo, John A. Paulson School of Engineering and Applied Sciences, Harvard University

David Weinberger, Berkman Klein Center for Internet and Society at Harvard University

Alexandra Wood, Berkman Klein Center for Internet and Society at Harvard University

Abstract

The ubiquity of systems using artificial intelligence or “AI” has brought increasing attention to how those systems should be regulated. The choice of how to regulate AI systems will require care. AI systems have the potential to synthesize large amounts of data, allowing for greater levels of personalization and precision than ever before—applications range from clinical decision support to autonomous driving and predictive policing. That said, our AIs continue to lag in common sense reasoning [McCarthy, 1960], and thus there exist legitimate concerns about the intentional and unintentional negative consequences of AI systems [Bostrom, 2003, Amodei et al., 2016, Sculley et al., 2014].

How can we take advantage of what AI systems have to offer, while also holding them accountable? In this work, we focus on one tool: explanation. Questions about a legal right to explanation from AI systems was recently debated in the EU General Data Protection Regulation [Goodman and Flaxman, 2016, Wachter et al., 2017a], and thus thinking carefully about when and how explanation from AI systems might improve accountability is timely. Good choices about when to demand explanation can help prevent negative consequences from AI systems, while poor choices may not only fail to hold AI systems accountable but also hamper the development of much-needed beneficial AI systems.

Below, we briefly review current societal, moral, and legal norms around explanation, and then focus on the different contexts under which explanation is currently required under the law. We find that there exists great variation around when explanation is demanded, but there also exist important consistencies: when demanding explanation from humans, what we typically want to know is whether and how certain input factors affected the final decision or outcome.

These consistencies allow us to list the technical considerations that must be considered if we desired AI systems that could provide kinds of explanations that are currently required of humans under the law. Contrary to popular wisdom of AI systems as indecipherable black boxes, we find that this level of explanation should generally be technically feasible but may sometimes be practically onerous—there are certain aspects of explanation that may be simple for humans to provide but challenging for AI systems, and vice versa. As an interdisciplinary team of legal scholars, computer scientists, and cognitive scientists, we recommend that for the present, AI systems can and should be held to a similar standard of explanation as humans currently are; in the future we may wish to hold an AI to a different standard.

1 Introduction

AI systems are currently used in applications ranging from automatic face-focus on cameras [Ray and Nicponski, 2005] and predictive policing [Wang et al., 2013] to segmenting MRI scans [Aibinu et al., 2008] and

language translation [Chand, 2016]. We expect that they will be soon be applied in safety-critical applications such as clinical decision support [Garg et al., 2005] and autonomous driving [Maurer et al., 2016]. That said, AI systems continue to be poor at common sense reasoning [McCarthy, 1960]. Thus, there exist legitimate concerns about the intentional and unintentional negative consequences of AI systems [Bostrom, 2003, Amodei et al., 2016, Sculley et al., 2014].

How can we take advantage of what AI systems have to offer, while also holding them accountable? To date, AI systems are only lightly regulated: it is assumed that the human user will use their common sense to make the final decision. However, even today we see many situations in which humans place too much trust in AI systems and make poor decisions—consider the number of car accidents due to incorrect GPS directions [Wolfe, February 17, 2014], or, at a larger scale, how incorrect modeling assumptions were at least partially responsible for the recent mortgage crisis [Donnelly and Embrechts, 2010]. As AI systems are used in more common and consequential contexts, there is increasing attention on whether and how they should be regulated. The question of how to hold AI systems accountable is important and subtle: poor choices may result in regulation that not only fails to truly improve accountability but also stifles the many beneficial applications of AI systems.

While there are many tools to increasing accountability in AI systems, we shall focus on one in this report: explanation. (We briefly discuss alternatives in Section 7.) By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute. The question of when and what kind of explanation might be required of AI systems is urgent: details about a potential “right to explanation” were debated in the most recent revision of the European Union’s General Data Protection Regulation (GDPR) [Goodman and Flaxman, 2016, Wachter et al., 2017a]. While the ultimate version of the GDPR only requires explanation in very limited contexts, we expect questions around AI and explanation to be important in future regulation of AI systems—and, as noted above, it is essential that such regulation is implemented thoughtfully. In particular, there exist concerns that the engineering challenges surrounding explanation from AI systems would stifle innovation; that explanations might force trade secrets to be revealed; and that explanation would come at the price of system accuracy or other performance objective.

In this document, we first examine what kinds questions legally-operative explanations must answer. We then look at how explanations are currently used by society and, more specifically, in our legal and regulatory systems. We find that while there is little consistency about when explanations are required, there is a fair amount of consistency in what the abstract form of an explanation needs to be. This property is very helpful for creating AI systems to provide explanation; in the latter half of this document, we describe technical considerations for designing AI systems to provide explanation while mitigating concerns about sacrificing prediction performance and divulging trade secrets. Under legally operative notions of explanations, AI systems are not indecipherable black-boxes; we can, and sometimes should, demand explanation from them. We also discuss the potential costs of requiring explanation from AI systems, situations in which explanation may not be appropriate, and finally other ways of holding AI systems accountable.

This document is a product of over a dozen meetings between legal scholars, computer scientists, and cognitive scientists. Together, we are experts on explanation in the law, on the creation of AI systems, and on the capabilities and limitations of human reasoning. This interdisciplinary team worked together to recommend what kinds of regulation on explanation might be both beneficial and feasible from AI systems.

2 What is an Explanation?

In the colloquial sense, any clarifying information can be an explanation. Thus, we can “explain” how an AI makes decision in the same sense that we can explain how gravity works or explain how to bake a cake: by laying out the rules the system follows without reference to any specific decision (or falling object, or cake). When we talk about an explanation for a decision, though, we generally mean the reasons or justifications for that particular outcome, rather than a description of the decision-making process in general. In this paper, when we use the term explanation, we shall mean a human-interpretable description of the process by which

a decision-maker took a particular set of inputs and reached a particular conclusion [Wachter et al., 2017a] (see Malgieri and Comandè [2017] for a discussion about legibility of algorithmic systems more broadly).

In addition to this formal definition of an explanation, an explanation must also have the correct type of content in order for it to be useful. As a governing principle for the content an explanation should contain, we offer the following: an explanation should permit an observer to determine the extent to which a particular input was determinative or influential on the output. Another way of formulating this principle is to say that an explanation should be able to answer at least one of the following questions:

What were the main factors in a decision? This is likely the most common understanding of an explanation for a decision. In many cases, society has prescribed a list of factors that must or must not be taken into account in a particular decision. For example, we many want to confirm that a child’s interests were taken into account in a custody determination, or that race was not taken into account in a criminal prosecution. A list of the factors that went into a decision, ideally ordered by significance, helps us regulate the use of particularly sensitive information.

Would changing a certain factor have changed the decision? Sometimes, what we want to know is not whether a factor was taken into account at all, but whether it was determinative. This is most helpful when a decision-maker has access to a piece of information that has both improper and proper uses, such as the consideration of race in college admissions. By looking at the effect of changing that information on the output and comparing it to our expectations, we can infer whether it was used correctly.

Why did two similar-looking cases get different decisions, or vice versa? Finally, we may want to know whether a specific factor was determinative in relation to another decision. This information is useful when we need to assess the consistency as well as the integrity of a decision-maker. For example, it would be proper for a bank to take income into account, and even treat it as dispositive, when deciding whether to grant a loan. However, we might not want a bank to rely on income to different degrees in apparently similar cases, as this could undermine the predictability and trustworthiness of the decision-making process.

3 Societal Norms Around Explanation

Before diving into the U.S. legal context, we discuss more broadly how we, as a society, find explanations are desirable in some circumstances but not others. In doing so, we lay the foundations for specific circumstances in which explanation are (or are not) currently demanded under the law (Section 4). When it comes to human decision-makers, we often want an explanation when someone makes a decision we do not understand or believe to be suboptimal [Leake, 1992]. For example, was the conclusion accidental or intentional? Was it caused by incorrect information or faulty reasoning? The answers to these questions permit us to weigh our trust in the decision-maker and to assign blame in case of a dispute.

However, society cannot demand an explanation for every decision, because explanations are not free. Generating them takes time and effort, thus reducing the time and effort available to spend on other, potentially more beneficial conduct. Therefore, the utility of explanations must be balanced against the cost of generating them. Consider the medical profession. A doctor who explained every diagnosis and treatment plan to another doctor might make fewer mistakes, but would also see fewer patients. And so, we required newly graduated doctors to explain their decisions to more senior colleagues, but we do not require explanation from more experienced doctors—as the risk of error decreases and the value of the doctor’s time increases, the cost-benefit analysis of generating explanations shifts.

In other circumstances, an explanation might obscure more information than it reveals—humans are notoriously inaccurate when providing post-hoc rationales for decisions [Nisbett and Wilson, 1977]—and even if an explanation is accurate, we cannot ensure that it will be used in a socially responsible way. Explanations can also change an individual’s judgment: the need to explain a decision can have both positive and negative effects on the decision-maker’s choices [Messier et al., 1992], and access to an explanation might

decrease observers' trust in some decisions [de Fine Licht, 2011]. Last but not least, social norms regarding individual autonomy weigh against demanding explanations for highly personal decisions.

What, then, are the circumstances in which the benefits of an explanation outweigh the costs? We find that there are three conditions that characterize situations in which society considers a decision-maker is obligated—morally, socially, or legally—to provide an explanation:

The decision must have been acted on in a way that has an impact on a person other than the decision maker. If a decision only impacts the decision-maker, social norms generally will not compel an explanation, as doing so would unnecessarily infringe upon the decision-maker's independence. For example, if an individual invests their own funds and suffers losses, there is no basis to demand that the investor disclose their strategy. But if an investor makes a decision that loses a client's money, the client may well be entitled to an explanation.

There must be value to knowing if the decision was made erroneously. Assuming the decision affects entities other than the decision-maker, society still will not demand an explanation unless the explanation can be acted on in some way. Under the law, this action usually corresponds to assigning a blame and providing compensation for injuries caused by past decisions. However, as noted in Wachter et al. [2017b], explanations can also be useful if they can positively change future decision-making. But if there is no recourse for the harm caused, then there is no justification for the cost of generating an explanation. For example, if a gambler wins a round of roulette, there is no reason to demand an explanation for the bet: there is no recourse for the casino and there is no benefit to knowing the gambler's strategy, as the situation is not repeatable.

There must be some reason to believe that an error has occurred (or will occur) in the decision-making process. We only demand explanations when some element of the decision-making process—the inputs, the output, or the context of the process—conflicts with our expectation of how the decision will or should be made:

- **Unreliable or inadequate inputs.** In some cases, belief that an error has occurred arises from our knowledge of the decision-maker's inputs. An input might be suspect because we believe it is logically irrelevant. For example, if a surgeon refuses to perform an operation because of the phase of the moon, society might well deem that an unreasonable reason to delay an important surgery [Margot, 2015]. An input might also be forbidden. Social norms in the U.S. dictate that certain features, such as race, gender, and sexual identity or orientation, should not be taken into account deciding a person's access to employment, housing, and other social goods. If we know that a decision-maker has access to irrelevant or forbidden information—or a proxy for such information—it adds to our suspicion that the decision was improper. Similarly, there are certain features that we think *must* be taken into account for particular decision: if a person is denied a loan, but we know that the lender never checked the person's credit report, we might suspect that the decision was made on incomplete information and, therefore, erroneous.
- **Inexplicable outcomes.** In other cases, belief that an error occurred comes from the output of the decision-making process, that is, the decision itself. If the same decision-maker renders different decisions for two apparently identical subjects, we might suspect that the decision was based on an unrelated feature, or even random. Likewise, if a decision-maker produces the same decision for two markedly different subjects, we might suspect that it failed to take into account a salient feature. Even a single output might defy our expectations to the degree that the most reasonable inference is that the decision-making process was flawed. If an autonomous vehicles suddenly veers off the road, despite there being no traffic or obstacles in sight, we could reasonably infer that an error occurred from that single observation.

- **Distrust in the integrity of the system.** Finally, we might demand an explanation for a decision even if the inputs and outputs appear proper because of the context in which the decision is made. This usually happens when a decision-maker is making highly consequential decisions and has the ability or incentive to do so in a way that is personally beneficial but socially harmful. For example, corporate directors may be tempted to make decisions that benefit themselves at the expense of their shareholders. Therefore, society may want corporate boards to explain their decisions, publicly and preemptively, even if the inputs and outputs of the decision appear proper [Hopt, 2011].

We observe that the question of when it is reasonable to demand an explanation is more complex than identifying the presence or absence of these three factors. Each of these three factors may be present in varying degree, and no single factor is dispositive. When a decision has resulted in a serious and plainly redressable injury, we might require less evidence of improper decision-making. Conversely, if there is a strong reason to suspect that a decision was improper, we might demand an explanation for even a relatively minor harm. Moreover, even where these three factors are absent, a decision-maker may want to voluntarily offer an explanation as a means of increasing trust in the decision-making process. To further demonstrate the complexity of determining when to requiring explanations, we now look at a concrete example: the U.S. legal system.

4 Explanations in the U.S. Legal System

The principles described in Section 3 describe the general circumstances in which we, as a society, desire explanation. We now consider how they are applied in existing laws governing human behavior. We confine our research to laws for two reasons. First, laws are concrete. Reasonable minds can and do differ about whether it is morally justifiable or socially desirable to demand an explanation in a given situation. Laws on the other hand are codified, and while one might argue whether a law is correct, at least we know what the law is. Second, the United States legal system maps well on to the three conditions from Section 3. The first two conditions—that the decision have an actual effect and that there is some benefit to obtaining an explanation—are embodied in the doctrine of standing within the constitutional injury, causation, and redressability requirements [Krent, 2001]. The third condition, reason to believe that an error occurred, corresponds to the general rule that the complaining party must allege some kind of mistake or wrongdoing before the other party is obligated to offer an explanation—in the legal system, this is called “meeting the burden of production” [Corpus Juris Secundum, c. 86 §101]. Indeed, at a high level, the anatomy of many civil cases involve the plaintiff presenting evidence of an erroneous decision, forcing the defendant to generate an innocent explanation or concede that an error occurred.

However, once we get beyond this high-level model of the legal system, we find significant variations in the demand for explanations under the law, including the role of the explanation, who is obligated to provide it, and what type or amount of evidence is needed to trigger that obligation. A few examples that highlight this variation follow:

- **Strict liability:** Strict liability is a form of legal liability that is imposed solely on the fact that the defendant caused an injury; there is no need to prove that the defendant acted wrongfully, intentionally, or even negligently. Accordingly, the defendant’s explanation for the decision to act in a certain way is irrelevant to the question of liability. Strict liability is usually based on risk allocation policies. For example, under U.S. product liability law, a person injured as a result of a poor product design decision can recover damages without reaching the question of *how* that decision was made. The intent of the strict product liability system is to place the burden of inspecting and testing products on manufacturers, who have the resources and expertise to do so, rather than consumers, who presumably do not [Owen and Davis, 2017, c. 1 §5:1].
- **Divorce:** Prior to 1969, married couples in the U.S. could only obtain a divorce by showing that one of the spouses committed some wrongful act such as abuse, adultery, or desertion—what are called “grounds for divorce.” Starting with California in 1969, changing social norms around around privacy

and autonomy, especially for women, led states to implement no-fault divorce laws, under which a couple can file for divorce without offering a specific explanation. Now, all states provide for no-fault divorce, and requiring a couple to explain their decision to separate is perceived as archaic [Guidice, 2011].

- **Discrimination:** In most discrimination cases, the plaintiff must provide some evidence that some decision made by the defendant—for example, the decision to extend a government benefit to the plaintiff—was intentionally biased before the defendant is required to present a competing explanation [Strauss, 1989]. But in certain circumstances, such as criminal jury selection, employment, or access to housing, statistical evidence that the outputs of a decision-making process disproportionately exclude a particular race or gender is enough to shift the burden of explanation on the decision-maker [Swift, 1995, Cummins and Isle, 2017]. This stems in part from the severity and prevalence of certain types of discrimination, but also a moral judgment about the repugnance of discriminating on certain characteristics.
- **Administrative decisions:** Administrative agencies are subject to different explanation requirements at different stages in their decision-making. When a new administrative policy is being adopted, the agency must provide a public explanation for the change [Corpus Juris Secundum, c. 73 §231]. But once the policies are in place, a particular agency decision is usually given deference, meaning that a court reviewing the decision will assume that the decision is correct absent countervailing evidence. Under the deferential standard, the agency only needs to show that the decision was not arbitrary or random [Corpus Juris Secundum, c. 73A §497]. Highly sensitive decisions, like national security related decisions, may be immune from any explanatory requirement at all.
- **Judges and juries:** Whether and how a particular judicial decision must be explained varies based on a number of factors, including the importance of the decision and the nature of the decision-maker. For example, a judge ruling on a motion to grant a hearing can generally do so with little or no explanation; the decision is highly discretionary. But a judge handing down a criminal sentence—one of the most important decisions a court can make—must provide an explanation so that the defendant can detect and challenge any impropriety or error [O’Hear, 2009]. On the other hand, a jury cannot be compelled to explain why it believed a certain witness or drew a certain inference, even though these decisions may have an enormous impact on the parties. One justification given for not demanding explanations from juries is that public accountability could bias jurors in favor of making popular but legally incorrect decisions; another is that opening jury decisions to challenges would weaken public confidence in the outcomes of trials and bog down the legal system [Landsman, 1999].

As the foregoing examples show, even in the relatively systematic and codified realm of the law, there are numerous factors that affect whether human decision-makers will be required to explain their decisions. These factors include the nature of the decision, the susceptibility of the decision-maker to outside influence, moral and social norms, the perceived costs and benefits of an explanation, and a degree of historical accident.

5 Implications for AI systems

With our current legal contexts in mind, we now turn to technical considerations for extracting explanation from AI systems. That is, how challenging would it be to create AI systems that provide the same kinds of explanation that are currently expected of humans, in the contexts that are currently expected of humans, under the law? Human decision-makers are obviously different from AI systems (see Section 6 for a comparison), but in this section we answer this question largely in the affirmative: for the most part, it *is* technically feasible to extract the kinds of explanations that are currently required of humans from AI systems.

Legally-Operative Explanations are Feasible. The main source of this feasibility arises from the fact that explanation is *distinct* from transparency. Explanation does not require knowing the flow of bits through

an AI system, no more than explanation from humans requires knowing the flow of signals through neurons (neither of which would be interpretable to a human!). Instead, explanation, as required under the law, as outlined in Section 2, is about answering how certain factors were used to come to the outcome in a specific situation. These core needs can be formalized by two technical ideas: *local explanation* and *local counterfactual faithfulness*.

Local Explanation. In the AI world, explanation for a specific decision, rather than an explanation of the system’s behavior overall, is known as local explanation [Ribeiro et al., 2016, Lei et al., 2016, Adler et al., 2016, Fong and Vedaldi, 2017, Selvaraju et al., 2016, Smilkov et al., 2017, Shrikumar et al., 2016, Kindermans et al., 2017, Ross et al., 2017, Singh et al., 2016]. AI systems are naturally designed to have their inputs varied, differentiated, and passed through many other kinds of computations—all in a reproducible and robust manner. It is already the case that AI systems are trained to have relatively simple decision boundaries to improve prediction accuracy, as we do not want tiny perturbations of the input changing the output in large and chaotic ways [Drucker and Le Cun, 1992, Murphy, 2012]. Thus, we can readily expect to answer the first question in Section 2—what were the important factors in a decision—by systematically probing the inputs to determine which have the greatest effect on the outcome. This explanation is *local* in the sense that the important factors may be different for different instances. For example, for one person, payment history may be the reason behind their loan denial, for another, insufficient income.

Counterfactual Faithfulness. The second property, counterfactual faithfulness, encodes the fact that we expect the explanation to be causal. Counterfactual faithfulness allows us to answer the remaining questions from Section 2: whether a certain factor determined the outcome, and related, what factor caused a difference in outcomes. For example, if a person was told that their income was the determining factor for their loan denial, and then their income increases, they might reasonably expect that the system would now deem them worthy of getting the loan. Importantly, however, we only expect that counterfactual faithfulness apply for related situations—we would not expect an explanation in a medical malpractice case regarding an elderly, frail patient to apply to a young oncology patient. However, we may expect it to still hold for a similar elderly, less frail patient. Recently Wachter et al. [2017b] also point out how counterfactuals are the cornerstone of what we need from explanation.

Importantly, both of these properties above can be satisfied *without* knowing the details of how the system came to its decision. For example, suppose that the legal question is whether race played an inappropriate role in a loan decision. One might then probe the AI system with variations of the original inputs changing only the race. If the outcomes were different, then one might reasonably argue that gender played a role in the decision. And if it turns out that race played an inappropriate role, that constitutes a legally sufficient explanation—no more information is needed under the law (although the company may internally choose decide to determine the next level of cause, e.g. bad training data vs. bad algorithm). This point is important because it mitigates concerns around trade secrets: explanation can be provided without revealing the internal contents of the system.

Explanation systems should be considered distinct from AI systems. We argue that regulation around explanation from AI systems should consider the explanation system as *distinct* from the AI system. Figure 1 depicts a schematic framework for explainable AI systems. The AI system itself is a (possibly proprietary) black-box that takes in some inputs and produces some predictions. The designer of the AI system likely wishes the predictions (\hat{y}) to match the real world (y). The designer of the *explanation system* must output a *human-interpretable* rule $e_x()$ that takes in the same input x and outputs a prediction \tilde{y} . To be locally faithful under counterfactual reasoning formally means that the predictions \tilde{y} and \hat{y} are the same under small perturbations of the input x .

This framework renders concepts such as local explanation and local counterfactual faithfulness readily quantifiable. For any input x , we can check whether the prediction made by the local explanation (\tilde{y}) is the same as the prediction made by the AI system (\hat{y}). We can also check whether these predictions remain consistent over small perturbations of x (e.g. changing the race). Thus, not only can we measure what proportion of the time an explanation system is faithful, but we can also identify the specific instances in which it is not. From a regulatory perspective, this opens the door to regulation that requires that an AI

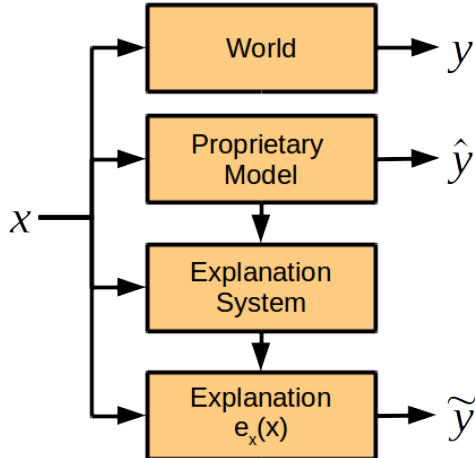


Figure 1: Diagram of a Framework for Explainable AI Systems.

system be explainable some proportion of the time or in certain kinds of contexts—rather than all the time. Loosening the explanation requirement in this way may allow for the AI system to use a much more complex logic for a few cases that really need it. More broadly, thinking of an explanation system as distinct from the original AI system also creates opportunities for industries that specialize in explanation systems.

There will exist challenges in mapping inputs and intermediate representations in AI systems to human-interpretable concepts. While the notion of how explanations are used under the law can be formalized computationally, there remains a key technical challenge of converting the inputs to an AI system—presumably some large collection of variables, such as pixel values—into human-interpretable terms such as age or gender. For example, self-driving cars may have multitudes of sensors, each with high-dimensional range and vision inputs; the human brain already converts its visual inputs into higher-level concepts such as trees or street signs. Clinical decision support systems may take in tens of thousands of variables about a patient’s diagnoses, drugs, procedures, and concepts extracted from the clinical notes; the human doctor has terms like sepsis or hypertension to describe constellations of these variables. While there do exist methods to map the high-dimensional inputs to an AI system to human-interpretable concepts, the process generally requires training the system with large amounts of data in which both the raw input and the associated concept are given.

As such, explanations from AI systems will be most straight-forward if the relevant terms are known in advance. In this case, the AI system can be trained to map its inputs to the relevant terms. For example, in the medical sphere, there are a number of algorithms for determining whether a patient has diabetes from a multitude of inputs [Newton et al., 2013]; recent work has identified ways to weigh the importance of much more general terms [Kim et al., 2017]. There will be some technical innovation required, but by and large we see relatively few difficulties for AI systems to provide the kinds of explanation that are currently required in the case where legislation or regulation makes it clear what terms may be asked for *ex ante*; there is also an established process for companies to adapt new standards as legislation and regulation change. That said, there are subtleties. While it is relatively straightforward to identify what inputs are correlated with certain terms, and verify whether predictions of terms are correlated with decisions, it will require some work to determine ways to test counterfactuals. For example, how can we show that a security system that uses images of a face as input does not discriminate against gender? One would need to consider an alternate face that was similar in every way except for gender.

Another subtlety is that, to create the required terms, the AI system will need access to potentially sensitive information. Currently, we often assume that if the human did not have access to a particular term, such as race, then it could not have been used in the decision. However, it is very easy for AI systems

to reconstruct sensitive terms from high-dimensional inputs. Data about shopping patterns can be used to identify term such as age, gender, and socio-economic status, as can data about healthcare utilization. Especially with AI systems, excluding a protected category does not mean that a proxy for that category is not being created. Thus, a corollary to the arguments above is that we must measure any terms that we wish to protect against, to be able to ensure that we are not generating proxies for them. Our legal system must allow them to be collected, and AI system designers should build ways to test whether systems are creating that term and using it inappropriately. Regulation must be put in place so that any protected terms collected by AI system designers are used only to ensure that the AI system is designed correctly, and not for other purposes within the organization. (It would be unfortunate, to say the least, if we can verify that an AI system is not discriminating against a protected term, only to find that a human decision-maker is accessing and combining the forbidden information with the AI system’s recommendation to make a final choice.)

The challenges increase if the relevant terms cannot be determined in advance. For example, in litigation scenarios, the list of relevant terms is generally only determined *ex post*. In such cases, AI systems may struggle; unlike humans, they cannot be asked to refine their explanations after the fact without additional training data. For example, we cannot identify what proxies there are for age in a data set if age itself has never been measured. For such situations, we first note that there is precedent for what to do in litigation scenarios when some information is not available, ranging from drawing inferences against the party that could have provided the information to imposing civil liability for unreasonable record-keeping practices [Nolte, 1994, Cicero, 1988]. Second, while not always possible, in many cases it may be possible to quickly train a proxy—especially if AI designers have designed the system to be updated—or have the parties mutually agree (perhaps via a third party) what are acceptable proxies. The parties may also agree to assessment via non-explanation-based tools.

In summary, to build AI systems that can provide explanation in terms of human-interpretable terms, we must both list those terms and allow the AI system access to examples to learn them. System designers should design systems to learn these human-interpretable terms, and also store data from each decision so that is possible to reconstruct and probe a decision post-hoc if needed. Policy makers should develop guidelines to ensure that the explanation system is being faithful to the original AI.

6 A Comparison of Human and AI Capability for Explanation

So far, we have argued that explanation from AI is technically feasible in many situations. However, there are obviously salient differences between AI systems and humans. Should this affect the extent to which AI explanations should be the subject of regulation? We begin with the position that, in general, AIs should be capable of providing an explanation in any situation where a human would be legally required to do so. This approach would prevent otherwise legally accountable decision-makers from “hiding” behind AI systems, while not requiring the developers of AI systems to spend resources or limit system performance simply to be able to generate legally unnecessary explanations.

That said, given the differences between human and AI processes, there may be situations in which it is possible to demand more from humans, and other situations in which it might be possible to hold AI systems to a higher standard of explanation. There are far too many factors that go into determining when an explanation should be legally required to analyze each of them with respect to both humans and AIs in this paper. At the most general level, though, we can categorize the factors that go into such a determination as either extrinsic or intrinsic to the decision-maker. Extrinsic factors—the significance of the decision, the relevant social norms, the extent to which an explanation will inform future action—are likely to be the same whether the decision-maker is a human or an AI system.

Intrinsic factors, though, may vary significantly between humans and AIs (see Table 1), and will likely be key in eventually determining where demands for human and AI explanations under the law should overlap and where they should diverge. One important difference between AIs and humans is the need to pre-plan explanations. We assume that humans will, in the course of making a decision, generate and store the information needed to explain that decision later if doing so becomes useful. A doctor who does not

explain the reasons for a diagnosis at the time it is made can nevertheless provide those reasons after the fact if, for example, diagnosis is incorrect and the doctor gets sued. A decision-maker might be required to create a record to aid in the subsequent generation of an explanation—to continue the prior example, many medical providers require doctors to annotate patient visits for this very reason, despite the fact that it takes extra time. However, requiring human decision-makers to document their decisions is the exception, not the norm. Therefore, the costs and benefits of generating an human explanation can be assessed at the time the explanation is requested.

In contrast, AI systems do not automatically store information about their decisions. Often, this feature is considered an advantage: unlike human decision-makers, AI systems can delete information to optimize their data storage and protect privacy. However, an AI system designed this way would not be able to generate *ex post* explanations the way a human can. Instead, whether resources should to be allocated to explanation generation becomes a question of system design. This is analogous to the question of whether a human decision-maker should be required to keep a record. The difference is that with an AI system this design question must *always* be addressed explicitly.

That said, AI systems can be designed to store their inputs, intermediate steps, and outputs exactly (although transparency may be required to verify this). Therefore, they do not suffer from the cognitive biases that make human explanations unreliable. Additionally, unlike humans, AI systems are not vulnerable to the social pressures that could alter their decision-making processes. Accordingly, there is no need to shield AI systems from generating explanations, for example, the way the law shields juries.

Table 1: Comparison of Human and AI Capabilities for Explanation

	<i>Human</i>	<i>AI</i>
<i>Strengths</i>	Can provide explanation post-hoc	Reproducible, no social pressure
<i>Weaknesses</i>	May be inaccurate and unreliable, feel social pressure	Requires up-front engineering, explicit taxonomies and storage

7 Alternatives to Explanation

Explanation is but one tool to hold AI systems accountable. In this section, we discuss the trade-offs associated with three core classes of tools: explanation, empirical evidence, and theoretical guarantees.

Explanation. In Section 5, we noted that an explanation system may struggle if a new factor is suddenly needed. In other cases, explanation may be possible but undesirable for other reasons: Designing a system to also provide explanation is a non-trivial engineering task, and thus requiring explanation all the time may create a financial burden that disadvantages smaller companies; if the decisions are low enough risk, we may not wish to require explanation. In some cases, one may have to make trade-offs between the proportion of time that explanation can be provided and the accuracy of the system; that is, by requiring explanation we might cause the system to reject a solution that cannot be reduced to a human-understandable set of factors. Obviously, both explanation and accuracy are useful for preventing errors, in different ways. If the overall number of errors is lower in a version of the AI system that does not provide explanation, then we might wish to only monitor the system to ensure that the errors are not targeting protected groups and the errors even out over an individual. Similar situations may occur even if the AI is not designed to reject solutions that fall below a threshold of explicability; the human responsible for implementing the solution may discard it in favor of a less optimal decision with a more appealing—or legally defensible—explanation. In either case, society would lose out on an optimal solution. Given that one of the purported benefits of AI decision-making is the ability to identify patterns that humans cannot, this would be counterproductive.

Empirical Evidence. Another tool for accountability is empirical evidence, that is measures of a system’s overall performance. Empirical evidence may justify (or implicate) a decision-making system by demonstrat-

ing the value (or harm) of the system, without providing an explanation for any given decision. For example, we might observe that an autonomous aircraft landing system has fewer safety incidents than human pilots, or that the use of a clinical diagnostic support tool reduces mortality. Questions of bias or discrimination can be ascertained statistically: for example, a loan approval system might demonstrate its bias by approving more loans for men than women when other factors are controlled for. In fact, in some cases statistical evidence is the only kind of justification that is possible; certain types of subtle errors or discrimination may only show up in aggregate. While empirical evidence is not unique to AI systems, AI systems, as digesters of data used in highly reproducible ways, are particularly well-suited to provide empirical evidence. However, such evidence, by its nature, cannot be used to assign blame or innocence surrounding a particular decision.

Theoretical Guarantees. In rarer situations, we might be able to provide theoretical guarantees about a system. For example, we trust our encryption systems because they are backed by proofs; neither explanation or evidence are required. Similarly, if there are certain agreed-upon schemes for voting and vote counting, then it may be possible to design a system that provably follows those processes. Likewise, a lottery is shown to be fair because it abides by some process, even though there is no possibility of fully explaining the generation of the pseudo-random numbers involved. Theoretical guarantees are a form of perfect accountability that only AI systems can provide, and ideally will provide more and more often in the long term; however, these guarantees require very cleanly specified contexts that often do not hold in real-world settings.

We emphasize that the trade-offs associated with all of these methods will shift as technologies change. For example, access to greater computational resources may reduce the computational burden associated with explanation, but enable even more features to be used, increasing the challenges associated with accurate summarization. New modes of sensing might allow us to better measure safety or bias, allowing us to rely more on empirical evidence, but they might also result in companies deciding to tackle even more ambitious, hard-to-formalize problems for which explanation might be the only available tool. We summarize considerations for choosing an accountability tool for AI systems in Table 2.

Table 2: Considerations for Approaches for Holding AIs Accountable

<i>Approach</i>	<i>Well-suited Contexts</i>	<i>Poorly-suited Contexts</i>
Theoretical Guarantees	Situations in which both the problem and the solution can be fully formalized (gold standard, for such cases)	Any situation that cannot be sufficiently formalized (most cases)
Statistical evidence	Problems in which outcomes can be completely formalized, and we take a strict liability view; problems where we can wait to see some negative outcomes happen so as to measure them	Situations where the objective cannot be fully formalized in advance
Explanation	Problems that are incompletely specified, where the objectives are not clear and inputs might be erroneous	Situations in which other forms of accountability are not possible

8 Recommendations

In the sections above, we have discussed the circumstances in which humans are required to provide explanation under the law, as well as what those explanations are expected to contain. We have also argued that

it should be technically feasible to create AI systems that provide the level of explanation that is currently required of humans. The question, of course, is whether we *should*. The fact of the matter is that AI systems are increasing in capability at an astounding rate, with optimization methods of black-box predictors that far exceed human capabilities. Making such quickly-evolving systems be able to provide explanation, while feasible, adds an additional amount of engineering effort that might disadvantage less-resourced companies because of the additional personnel hours and computational resources required; these barriers may in turn result in companies employing suboptimal but easily-explained models.

Thus, just as with requirements around human explanation, we will need to think about why and when explanations are useful enough to outweigh the cost. Requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable but suboptimal outcomes. For example, the overhead of forcing a toaster to explain why it thinks the bread is ready might prevent a company from implementing a smart toasting feature—either due to the engineering challenges or concerns about legal ramifications. On the other hand, we may be willing to accept the monetary cost of an explainable but slightly less accurate loan approval system for the societal benefit of being able to verify that it is nondiscriminatory. As discussed in Section 3, there are societal norms around when we need explanation, and these norms should be applied to AI systems as well.

For now, we posit that demanding explanation from AI systems in such cases is not so onerous that we should ask of our AI systems what we ask of humans. Doing so avoids AI systems from getting a “free pass” to avoid the kinds of scrutiny that may come to humans, and also avoids asking so much of AI systems that it would hamper innovation and progress. Even this modest step will have its challenges, and as they are resolved, we will gain a better sense of whether and where demands for explanation should be different between AI systems and humans. As we have little data to determine the actual costs of requiring AI systems to generate explanations, the role of explanation in ensuring accountability must also be re-evaluated from time to time, to adapt with the ever-changing technology landscape.

Acknowledgements The BKC Working Group on Interpretability acknowledges Elena Goldstein, Jeffrey Fossett, and Sam Daitzman for helping organize our meetings. We also are indebted to countless conversations with our colleagues, who helped question and refine the ideas presented in this work.

References

- John McCarthy. *Programs with common sense*. RLE and MIT Computation Center, 1960.
- Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pages 277–284, 2003.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- D Sculley, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. Machine learning: The high-interest credit card of technical debt. 2014.
- Bryce Goodman and Seth Flaxman. EU regulations on algorithmic decision-making and a ‘right to explanation’. In *ICML workshop on human interpretability in machine learning (WHI 2016)*, New York, NY. <http://arxiv.org/abs/1606.08813> v1, 2016.
- Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2): 76–99, 2017a.
- Lawrence A Ray and Henry Nicponski. Face detecting camera and method, September 6 2005. US Patent 6,940,545.

- Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer, 2013.
- Abiodun M Aibinu, Momoh JE Salami, Amir A Shafie, and Athaur Rahman Najeeb. Mri reconstruction using discrete fourier transform: a tutorial. *World Academy of Science, Engineering and Technology*, 42: 179, 2008.
- Sunita Chand. Empirical survey of machine translation tools. In *Research in Computational Intelligence and Communication Networks (ICRCIN), 2016 Second International Conference on*, pages 181–185. IEEE, 2016.
- Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, PJ Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, 293(10):1223–1238, 2005.
- Markus Maurer, J Christian Gerdes, Barbara Lenz, and Hermann Winner. *Autonomous driving: technical, legal and social aspects*. Springer Publishing Company, Incorporated, 2016.
- Sarah Wolfe. Driving into the ocean and 8 other spectacular fails as gps turns 25. *Public Radio International*, February 17, 2014.
- Catherine Donnelly and Paul Embrechts. The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *ASTIN Bulletin: The Journal of the IAA*, 40(1):1–33, 2010.
- Gianclaudio Malgieri and Giovanni Comandè. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, 2017.
- David Leake. *Evaluating Explanations: A Content Theory*. New York: Psychology Press, 1992.
- Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.
- William F. Messier, Jr, William C. Quilliam, D. E. Hirst, and Don Craig. The effect of accountability on judgment: Development of hypotheses for auditing; discussions; reply. *Auditing*, 11:123, 1992. URL <http://search.proquest.com.ezp-prod1.hul.harvard.edu/docview/216730107?accountid=11311>.
- Jenny de Fine Licht. Do we really want to know? the potentially negative effect of transparency in decision making on perceived legitimacy. *Scandinavian Political Studies*, 34(3):183–201, 2011. ISSN 1467-9477. doi: 10.1111/j.1467-9477.2011.00268.x. URL <http://dx.doi.org/10.1111/j.1467-9477.2011.00268.x>.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *arXiv preprint arXiv:1711.00399*, 2017b.
- Jean-Luc Margot. No evidence of purported lunar effect on hospital admission rates or birth rates. *Nursing research*, 64(3):168, 2015.
- Klaus J Hopt. Comparative corporate governance: The state of the art and international regulation. *The American Journal of Comparative Law*, 59(1):1–73, 2011.
- Harold Krent. Laidlaw: Redressing the law of redressability. *Duke Environmental Law and Policy Forum*, 12(1):85–117, 2001.
- Corpus Juris Secundum.
- David G. Owen and Mary J. Davis. *Owen & Davis on Product Liability*, 4th edition, 2017.

- Lauren Guidice. New york and divorce: Finding fault in a no fault system. *Journal of Law and Policy*, 19(2):787–862, 2011.
- David A Strauss. Discriminatory intent and the taming of brown. *The University of Chicago Law Review*, 56(3):935–1015, 1989.
- Joel H. Swift. The unconventional equal protection jurisprudence of jury selection. *Northern Illinois University Law Review*, 16:295–341, 1995.
- Justin D. Cummins and Belle Isle. Toward systemic equality: Reinvigorating a progressive application of the disparate impact doctrine. *Mitchell Hamline Law Review*, 43(1):102–139, 2017.
- Michael M O’Hear. Appellate review of sentence explanations: Learning from the wisconsin and federal experiences. *Marquette Law Review*, 93:751–794, 2009.
- Stephan Landsman. The civil jury in America. *Law and Contemporary Problems*, 62(2):285–304, 1999.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. In *KDD*, 2016.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.
- Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1–10. IEEE, 2016.
- Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.
- Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2016.
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Interpretable deep learning by propagating activation differences. *ICML*, 2016.
- Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, and Sven Dähne. Patternet and patternlrp—improving the interpretability of neural networks. *arXiv preprint arXiv:1705.05598*, 2017.
- Andrew Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence*, 2017.
- Sameer Singh, Marco Tulio Ribeiro, and Carlos Guestrin. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.
- Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.
- Kevin P Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

- Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. *Journal of the American Medical Informatics Association*, 20(e1):e147–e154, 2013.
- Been Kim, Justin Gilmer, Ulfar Erlingsson, Fernanda Viegas, and Martin Wattenberg. Tcav: Relative concept importance testing with linear concept activation vectors. 2017.
- Steffen Nolte. The spoliation tort: An approach to underlying principles. . *Mary's LJ*, 26:351, 1994.
- Michael Cicero. Drug testing of federal government employees: Is harm resulting from negligent record maintenance actionable. *U. Chi. Legal F.*, page 239, 1988.

An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System

November 9, 2017

The following open letter—signed by Harvard and MIT-based faculty, staff, and researchers Chelsea Barabas, Christopher Bavitz, Ryan Budish, Karthik Dinakar, Cynthia Dwork, Urs Gasser, Kira Hessekiel, Joichi Ito, Ronald L. Rivest, Madars Virza, and Jonathan Zittrain—is directed to the Massachusetts Legislature to inform its consideration of risk assessment tools as part of ongoing criminal justice reform efforts in the Commonwealth.

Dear Members of the Massachusetts Legislature:

We write to you in our individual capacities¹ regarding the proposed introduction of actuarial risk assessment (“RA”) tools in the Commonwealth’s criminal justice system. As you are no doubt aware, Senate Bill 2185²—passed by the Massachusetts Senate on October 27, 2017—mandates implementation of RA tools in the pretrial stage of criminal proceedings. Specifically:

- Section 182 of the bill would amend Massachusetts General Laws chapter 276 to include the following new Section 58E(a):

Subject to appropriation, pretrial services shall create or choose a risk assessment tool that analyzes risk factors to produce a risk

assessment classification for a defendant that will aid the judicial officer in determining pretrial release or detention under sections 58 to 58C, inclusive. Any such tool shall be tested and validated in the commonwealth to identify and eliminate unintended economic, race, gender or other bias.³

- Amendment 146 (which was adopted) would add language to chapter 276 requiring that “[a]ggregate data that concerns pretrial services shall be available to the public in a form that does not allow an individual to be identified.”⁴
- Amendment 147 (which was also adopted) would add language providing that “[i]nformation about any risk assessment tool, the risk factors it analyzes, the data on which analysis of risk factors is based, the nature and mechanics of any validation process, and the results of any audits or tests to identify and eliminate bias, shall be a public record and subject to discovery.”⁵

As researchers with a strong interest in algorithms and fairness, we recognize that RA tools may have a place in the criminal justice system. In some cases, and by some measures, use of RA tools may promote outcomes better than the status quo. That said, we are concerned that the Senate Bill’s implementation of RA tools is cursory and does not fully address the complex and nuanced issues implicated by actuarial risk assessments.

¹ For purposes of identification, we note that all signatories to this letter are Harvard- and MIT-based faculty and researchers whose work touches on issues relating to algorithms. Most of the undersigned are involved in a research initiative underway at the MIT Media Lab and Harvard University’s Berkman Klein Center for Internet & Society that seeks to examine ethics and governance concerns arising from the use of artificial intelligence, algorithms, and machine learning technologies. See AI Ethics and Governance, MIT Media Lab, <https://www.media.mit.edu/projects/ai-ethics-and-governance/overview/> (last visited Oct. 28, 2017); Ethics and Governance of Artificial Intelligence, Berkman Klein Ctr. for Internet & Society, <https://cyber.harvard.edu/research/ai> (last visited Oct. 28, 2017).

² S.B. 2185, 190th Gen. Court (Mass. 2017), available at <https://malegislature.gov/Bills/190/S2185.pdf> (last visited Nov. 2, 2017).

³ Id. § 182, 1808–12.

⁴ Id. Amendment 146, ID: S2185–146-R1, available at <https://malegislature.gov/Bills/GetAmendmentContent/190/S2185/146/Senate/Preview> (last visited Oct. 29, 2017).

⁵ Id. Amendment 147, ID: S2185–147 (2017), available at <https://malegislature.gov/Bills/GetAmendmentContent/190/S2185/147/Senate/Preview> (last visited Oct. 29, 2017).

Find the original version of this letter on *Medium*
<https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9>

The success or failure of pretrial risk assessments in the Commonwealth will depend on the details of their design and implementation. Such design and implementation must be: (a) based on research and data; (b) accompanied (and driven) by clear and unambiguous policy goals; and (c) governed by principles of transparency, fairness, and rigorous evaluation.

As the Massachusetts House considers criminal justice reform legislation, and as both houses of the Legislature seek to reconcile their bills, we urge the Commonwealth to engage in significant study and policy development in this area. That study and policy development should ideally take place before the Legislature issues a mandate regarding adoption of risk assessment tools or, at the very least, before any particular tool is developed, procured, and/or implemented. As described herein, we submit that thoughtful deliberation is particularly important in five critical areas.

(1) The Commonwealth should take steps to mitigate the risk of amplifying bias in the justice system.

Research shows the potential for risk assessment tools to perpetuate racial and gender bias.⁶ Researchers have proposed multiple “fairness criteria” to mitigate this bias statistically.⁷ But there remain intrinsic tradeoffs between fairness and accuracy that are mathematically impossible for any RA tool to overcome. Senate Bill 2185 in-

cludes a single sentence on eliminating bias; we submit that this issue deserves far more consideration and deliberation.

Before implementing any RA tool, the Commonwealth should consider developing specific criteria along the following lines:

- (a) The Commonwealth should develop fairness criteria that mitigate the risk of an RA tool exacerbating bias on the basis of race, gender, and other protected classes.
- (b) The Commonwealth should craft rules and guidelines for identifying and ethically handling “proxy variables” (which correlate with race, gender, and other protected characteristics in any RA tool) and addressing other means by which such characteristics may be inferred from ostensibly neutral data. Notably in this regard, the state of California—which moved toward use of pretrial risk assessment tools relatively early—is now actively considering legislation to eliminate housing status and employment status from risk assessments, because these variables are strong proxies for race and class.⁸ If passed, such legislation would require counties to alter and adapt the patchwork of individual pretrial risk assessment tools in use across that state.⁹ We submit that the Commonwealth might learn from this example by putting in work upfront to fully understand bias and address proxies, rather than moving forward with implementation and specifying change at a later date.

6 See, e.g., Alexandra Chouldechova, Fair prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, arXiv:1703.00056 (submitted on Feb. 28, 2017), available at <https://arxiv.org/abs/1703.00056> (last visited Oct. 28, 2017); Devlin Barrett, Holder Cautions on Risk of Bias in Big Data Use in Criminal Justice, Wall St. J., Aug. 1, 2014, <https://www.wsj.com/articles/u-s-attorney-general-cautions-on-risk-of-bias-in-big-data-use-in-criminal-justice-1406916606> (last visited Oct. 28, 2017); Michael Tonry, Legal and Ethical Issues in the Prediction of Recidivism, 26 Fed. Sentencing Reporter 167, 173 (2014).

7 Richard Berk et al., Fairness in Criminal Justice Risk Assessments: The State of the Art, arXiv:1703.09207 (submitted on Mar. 27, 2017, last rev. 28 May 2017), available at <https://arxiv.org/abs/1703.09207> (last visited Oct. 28, 2017).

8 See Sonja B. Starr, Evidence-Based Sentencing and the Scientific Rationalization of Discrimination, 66 Stan. L. Rev. 803 (2014), available at <https://www.stanfordlawreview.org/print/article/evidence-based-sentencing-and-the-scientific-rationalization-of-discrimination/> (last visited Nov. 2, 2017).

9 See S.B. 10, 2017–2018 Reg. Sess. (Cal. 2017), available at http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201720180SB10 (last visited Nov. 1, 2017).

(c) The Commonwealth should create guidelines that govern data used in the development and validation of RA tools, to ensure tools deployed in Massachusetts are appropriately well-tailored to local populations and demographic structures.

(2) The Commonwealth should clarify procedures for validation and evaluation of risk assessment tools.

Research has shown that RA tools must be evaluated regularly and repeatedly to ensure their validity over time.¹⁰ In providing for adoption and use of risk assessments, the Commonwealth should take the opportunity to establish baselines concerning such review and evaluation. In particular, we urge the development of the following kinds of specifications:

(a) The Commonwealth should require mandatory, jurisdiction-by-jurisdiction validation checks, including rigorous comparison of a given tool’s predictions to observed results (such as re-conviction and failure to appear in court).

(b) The Commonwealth should insist that RA tools are tested on a regular basis to measure the disparate impact of tool error rates by race, gender, and other protected classes and should ensure that researchers have access to data and algorithms necessary to support robust testing.

(c) The Commonwealth should develop processes to promote regular (e.g., bi-annual) external oversight of validation checks of RA tools by an independent group—possibly a standing commission—which includes perspectives of statisticians, criminologists, and pretrial and probation service workers specific to the relevant jurisdiction.

(3) The Commonwealth should promulgate procedures for effective deployment of risk assessment tools.

Risk assessment tools employ statistical methods to produce risk scores. Representatives of the court system (usually, judges) use those numerical scores as one input in their pretrial decision-making processes, in the context of applicable legal standards. Use of an RA tool in a given case may involve a combination of statistical methods, fact determinations, and policy considerations. It is vital that all stakeholders in the pretrial pipeline be trained to accurately interpret and understand RA tools and the meaning (and limitations) of the risk assessment scores they produce.

By way of example, the classification of a risk category applicable to a particular criminal defendant with respect to a given risk score (e.g., high risk, medium risk, or low risk) is a matter of policy, not math. Tying the definition of terms like “high risk” to scores that are the products of RA tools can influence both: (a) decision-making by prosecutors, defendants, and judges in a pretrial setting (who may place undue emphasis on numerical scores generated by computers); and (b) public perception of the specific outcomes of RA tools. It is essential that the Commonwealth make clear how those risk scores are generated and what they purport to predict.

In this regard, we suggest the following:

(a) The Commonwealth should mandate continual training processes for all system actors to ensure consistency and reliability of risk score characterizations, irrespective of race, gender and other immutable characteristics.

¹⁰ See Risk and Needs Assessment and Race in the Criminal Justice System, Justice Ctr., Council State Gov’ts (May 31, 2016), <https://csgjusticecenter.org/reentry/posts/risk-and-needs-assessment-and-race-in-the-criminal-justice-system/> (last visited Oct. 28, 2017).

(b) The Commonwealth should require timely and transparent record-keeping practices that enable the auditing and adjustment of RA classifications over time.

(c) The Commonwealth should dictate a consistent decision-making framework to support appropriate interpretation of risk assessment predictions by all actors in the pretrial system. This framework should be regularly updated to reflect ongoing research about what specific conditions (i.e. electronic monitoring, weekly supervision meetings, etc.) have been empirically tested and proven to lower specific types of risk.

(d) The Commonwealth should provide adequate funding and resources for the formation and operation of an independent pretrial service agency that stands separate from other entities in the criminal justice system (such as probation offices and correctional departments). This agency will deal with the increased supervision caseload of individuals who are released prior to their trial date.

(e) The Commonwealth must ensure that updates to RA tools are accompanied by a detailed articulation of new intended risk characterizations.

(4) The Commonwealth should ensure that RA tools adequately distinguish among the types of risks being assessed.

A variety of risks may be relevant to a pre-trial determination such as bail. These risks may include (for example) the risk that a defendant will fail to appear for a hearing; the risk that a defendant will flee the jurisdiction; and the risk that defendant will engage in new criminal activity. Each of these risks may require different assessments, based on different factors, and each may need to be separately considered and weighed in

accordance with applicable legal standards in the context of a given pretrial decision.

Despite this complexity, most pretrial RA tools do not adequately differentiate among types of risks they purport to predict. An individual may be assigned a score indicating high risk in one category but not another, and the output report may not delineate this distinction. This can have significant implications for pretrial release decisions. A high risk of failure to appear in court due to mental health issues is not the same as a high risk that a defendant will commit a violent crime while awaiting trial. We urge the Legislature to ensure that RA tools adopted in the Commonwealth adequately differentiate among types of risks being assessed, so that courts can effectively identify appropriate conditions to place on defendants for release.

(5) The Commonwealth should give careful consideration to the process of developing or procuring RA tools, fully exploring the possibility of developing tools in-house, and establishing basic requirements for any tools developed by private vendors.

When a government entity seeks to adopt and implement any technological tool, it can do so in one of two ways. First, it can develop the tool on its own (relying on government personnel and/or outside developers). Second, it can purchase or license existing technology from a private outside vendor. In this regard, we submit that all of the factors identified in this letter should be considered by the Commonwealth with an eye toward informing two key decisions:

(a) a decision about whether Massachusetts should develop new risk assessment tools or procure existing ones; and

(b) establishing and enforcing concrete procurement criteria in the event the Commonwealth chooses to buy or license existing technology.

To the first point (re: whether to develop new tools or procure existing ones)—it is worth being mindful of cautionary tales such as the experience of local jurisdictions that sought to upgrade their voting infrastructures and implement electronic voting in the wake of the disputed 2000 United States presidential election.¹¹ Nearly twenty years later, many municipalities find themselves bound by undesirable contracts with a handful of outside vendors that offer unreliable voting machines and tallying services. Some of these vendors assert intellectual property protections in ways that complicate effective audits of the machines' accuracy and integrity.¹² Dissatisfaction with vendors is rarely sufficient to occasion a change in course, because of sunk costs and the burdens of reworking locked-in procedures. The Commonwealth must strive to avoid a structural repeat of governments' regrets around proprietary private voting infrastructure. There are strong arguments that the development of risk assessment tools for the justice system should be undertaken publicly rather than privately, that results should be shareable across jurisdictions, and that outcomes should be available for interrogation by the public at large.

To the second point (re: criteria for procurement)—we are hopeful that this document can serve as the basis for a roadmap toward develop-

ment of comprehensive procurement guidelines in the event that the Commonwealth decides to buy or license existing tools developed by private vendors rather than developing its own tools. Stated simply, procurement decisions cannot be based solely on considerations of cost or efficiency and must be driven by principles of transparency, accountability, and fairness. Those principles must be codified to ensure that the Commonwealth and its citizens leverage their purchasing power with vendors to understand what tools are being procured and ensure those tools operate fairly. Private vendors may raise concerns about scrutiny of their technologies and the algorithms they employ given proprietary business considerations. But, the Commonwealth must balance those private pecuniary interests against the overwhelming public interest in ensuring our criminal justice system satisfies fundamental notions of due process. The transparency measures described in Amendment 147 are welcome additions to the Senate Bill, and we urge consideration of additional measures that support fully-informed decision-making on this important issue.¹³

In conclusion, decisions around confinement and punishment are among the most consequential and serious that a government can make. They are non-delegable, and any technological aids that are not transparent, auditable, and improvable by the state cannot be deployed in the Commonwealth. Massachusetts has wisely avoided jumping rapidly into the use of RA tools. It is now in a position to consider them with the benefit of lessons

11 See, e.g., Andrew W. Appel et al., *The New Jersey Voting-Machine Lawsuit and the AVC Advantage DRE Voting Machine*, in *EVT/WOTE'09: Electronic Voting Technology Workshop / Workshop on Trustworthy Elections* (2009), available at <https://www.cs.princeton.edu/~appel/papers/appel-evt09.pdf> (last visited Nov. 2, 2017).

12 See, e.g., Alex Halderman, *How to Hack an Election in 7 Minutes*, *Politico* (Aug. 6, 2016), <https://www.politico.com/magazine/story/2016/08/2016-elections-russia-hack-how-to-hack-an-election-in-seven-minutes-214144> (last visited Oct. 28, 2017); David S. Levine, *Can We Trust Voting Machines?*, *Slate* (Oct. 24, 2012), www.slate.com/articles/technology/future_tense/2012/10/trade_secret_law_makes_it_impossible_to_independently_verify_that_voting.html (last visited Oct. 28, 2017).

13 By way of example, a recently proposed New York City Council Local Law would amend the administrative code of the City of New York to require agencies that use algorithms in certain contexts to both: (a) publish the source code used for such processing; and (b) accept user-submitted data sets that can be processed by the agencies' algorithms and provide the outputs to the user. See *Introduction N°1696-2017*, N.Y.C. Council (2017), available at <http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0> (last visited Oct. 28, 2017).

from jurisdictions that have gone first. We submit that—given that the potential benefits and dangers of pretrial RA tools rest on the details of tool development, oversight, and training, informed by clear policy goals—it is imperative that laws and regulations governing the introduction of pretrial RA tools be clear, concrete, specific, and data-driven. We are happy to assist in this effort.

Respectfully submitted,

Chelsea Barabas
Research Scientist,
MIT Media Laboratory

Christopher Bavitz
WilmerHale Clinical Professor of Law,
Harvard Law School

Ryan Budish
Assistant Research Director
Berkman Klein Center for Internet & Society

Karthik Dinakar
Research Scientist,
MIT Media Laboratory

Cynthia Dwork
Gordon McKay Professor of Computer Science,
Harvard School of Engineering and Applied Sciences
Radcliffe Alumnae Professor,
Radcliffe Institute for Advanced Study

Urs Gasser
Professor of Practice,
Harvard Law School

Kira Hessekiel
Project Coordinator,
Berkman Klein Center for Internet & Society

Joichi Ito
Director,
MIT Media Laboratory

Ronald L. Rivest
MIT Institute Professor

Madars Virza
Research Scientist,
MIT Media Laboratory

Jonathan Zittrain
George Bemis Professor of International Law,
Harvard Law School and Harvard Kennedy School
Professor of Computer Science,
Harvard School of Engineering and Applied Sciences

Appendix: Additional Readings

Below, we provide a selection of recent scholarship, analysis, and news coverage offering additional details about specific topics of potential interest to state Attorneys General seeking to better understand artificial intelligence. These readings are intended to be an introduction, and we encourage interested AGs to seek out the most recent coverage and to reach out to the Berkman Klein Center and this workshop's participants.

The following topics are covered below:

- General AI Mechanics
- Algorithmic Bias and Accountability
- Consumer Protection
- Criminal Justice
- Employment/Labor
- Privacy
- Duty of Explainability
- Government Use of AI
- Antitrust
- Broader Impacts of AI and Proposed Regulation

General AI Mechanics

Artificial Intelligence is a broad term that encompass various types of technology. The following pieces are good starting points for thinking about AI:

This video from DARPA describes three waves of Artificial Intelligence. The first is called Hand-crafted Knowledge in which defined rules were programmed to carry out specific tasks, like winning a game of chess or filing taxes. The video calls the current wave of AI technology, Statistical Learning. In this wave, systems are trained on large amounts of data to carry out tasks like voice and facial recognition. Finally, the video describes the the next wave, Contextual Adaptation wherein systems will be able themselves to build underlying explanatory models for describing real world phenomenon. DARPA Perspective on AI, Defense Advanced Research Projects Agency <https://www.darpa.mil/about-us/darpa-perspective-on-ai>.

This McKinsey report details five major limitations of Artificial Intelligence: data labeling (i.e. cleaning up raw data and organizing it for AI use), obtaining massive data training sets, explainability, generalizing the learning, and bias within the data/algorithms. Michael Chui, James Manyika, Medhi Miremadi, “What AI Can and Can’t Do (Yet) for Your Business,” McKinsey Quarterly (January 2018) <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business>.

This article provides a more technical overview of some of the key tools used by AI. “Artificial Intelligence Techniques Explained,” Deloitte. (April 2017) <https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-2-artificial-intelligence-techniques-explained.html>.

This article from the Berkman Center provides an overview of how the legal system may have to adjust in order to adequately address challenges raised by AI. Urs Gasser, “AI and the Law: Setting the Stage,” Medium. (June 2017) <https://medium.com/berkman-klein-center/ai-and-the-law-setting-the-stage-48516fda1b11>.

Criminal Justice

Algorithms are increasingly prevalent within our criminal justice system. Decisions on pretrial release, sentencing, and parole often consider the risk assessment scores produced by computer programs. These programs can look at factors ranging from the defendant’s prior criminal history to housing status to the defendant’s gender. Similar algorithms have also been used in predictive policing, directing law enforcement resources based on the assessed likelihood of future criminal activity.

Because these risk assessment algorithms are protected as trade secrets, neither courts nor defendants are given the chance to examine the underlying methodology. This lack of methodological transparency not only raises due process concerns, but has also prevented rigorous study of these programs and how they may perpetuate bias within the criminal justice system.

Risk Assessment/Sentencing

In one of the few studies of a risk assessment scoring program, ProPublica found that the scores were “remarkably unreliable in forecasting violent crimes,” and almost twice as likely to mistakenly predict general recidivism for black defendants than for white defendants. Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, “Machine Bias,” ProPublica (May 23, 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

This Brookings piece challenged the methodological issues of the ProPublica study. Jennifer L. Doleac, “Are Criminal Risk Assessment Scores Racist?,” Brookings (August 22, 2016) <https://www.brookings.edu/blog/up-front/2016/08/22/are-criminal-risk-assessment-scores-racist/>

This letter to the Massachusetts legislature highlights best practices for utilizing a risk assessment software in pretrial criminal proceedings. Chelsea Barabas, Christopher Bavitz, Ryan Budish, Karthik Dinakar, Cynthia Dwork, Urs Gasser, Kira Hessekziel, Joichi Ito, Ronald L. Rivest, Madars Virza, and Jonathan Zittrain, “An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System,” [Medium.com](https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9) (November 9, 2017) <https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9>

This law review case comment looks at the Wisconsin Supreme Court’s recent rejection of a due process challenge to the use of the COMPAS risk assessment tool for sentencing decisions. “State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessment in Scoring,” 130 Harv. L. Rev. 1530 (2017) <https://harvardlawreview.org/2017/03/state-v-loomis/>

This article provides a deep dive into Kentucky’s use of risk assessment tools in bail setting and discusses why more research is needed in order for such tools to have a significant impact. Megan T. Stevenson, “Assessing Risk Assessment in Action.” *Minnesota Law Review*, Vol. 103, (February 9, 2018) Forthcoming. Available at SSRN: <https://ssrn.com/abstract=3016088> or <http://dx.doi.org/10.2139/ssrn.3016088>

Predictive Policing

This paper, by academic and law enforcement advocates of predictive policing, discusses a study conducted with the Los Angeles Police Department that found an algorithm was significantly more effective in predicting crime hotspots. G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, “Randomized Controlled Field Trials of Predictive Policing.” *Journal of the American Statistical Association* (October 7, 2015) [https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2015.1077710?journalCode=uasa20#WuM_C8jRVPY]

This news article discusses the controversy over Los Angeles’ uses of predictive policing. Justin Jouvenal, “Police are using software to predict crime. Is it a ‘holy grail’ or biased against minorities?” (November 17, 2016) [https://www.washingtonpost.com/local/public-safety/police-are-using-software-to-predict-crime-is-it-a-holy-grail-or-biased-against-minorities/2016/11/17/525a6649-0472-440a-aae1-b283aa8e5de8_story.html?utm_term=.208c38063be4]

This article provides a skeptical overview of New Orleans’s experiment in using large amounts of citizen data to target law enforcement interventions; the article also discusses the controversy that arose regarding the program’s secrecy. Ali Winston, “Palantir Has Secretly Been Using New Orleans to Test its Predictive Policing Technology.” *The Verge*. (Feb 27, 2018) [<https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd>]

Consumer Protection

Artificial intelligence is being deployed across industries, redefining how companies advertise, design and deliver their products and services to consumers. In the consumer financial services sector, AI is often heralded as a way to more accurately measure creditworthiness than relying on traditional indicia such as FICO scores or income.

This article how AI allows lenders to rely on “alternative data” to assess a potential borrower’s credit risk. Such alternative data include daily location patterns, use of punctuation in text messages, or the proportion of phone contacts who have last names. The article also highlights how the use of nontraditional indicators can perpetuate biases. Charles Lane, “Will Using Artificial Intelligence To Make Loans Trade One Kind Of Bias For Another?” *NPR* (March 31, 2017) [<https://www.npr.org/sections/alltechconsidered/2017/03/31/521946210/will-using-artificial-intelligence-to-make-loans-trade-one-kind-of-bias-for-anot>]

This article highlights concerns with using algorithms to make consumer financial service decisions, including discrimination against populations who may not use computers and mobile devices, privacy violations, and appealing negative decisions. Penny Crosman, “Before AI Runs Amok, Banks Have Some Hard Decisions to Make,” *American Banker* (August 30, 2016) [<https://www.american-banker.com/news/before-ai-runs-amok-banks-have-some-hard-decisions-to-make>]

Autonomous Vehicles

Self-driving cars have been described as “the mother of all AI projects.” Not only are the technical challenges difficult, the stakes of making mistakes are higher than algorithms that recommend different products or a new show. Autonomous vehicles can also differ from other algorithmic technology because they can directly impact third parties (i.e. other drivers and pedestrians). Even as self-driving cars become more readily available, there are important questions about liability and consumer safety that have yet to be addressed.

This Department of Transportation website provides a useful overview on autonomous vehicles. Autonomous National Highway Traffic Safety Administration, Resources on Autonomous Vehicles <https://www.nhtsa.gov/technology-innovation>

Should an autonomous vehicle be programmed to drive off the road possibly killing its driver, rather than hitting and possibly killing several pedestrians? This TED talk addresses the ethical problems when regulating self-driving cars. Iyad Rahwan, “What Moral Decisions Should Driverless Cars Make?,” TEDxCambridge (September 2016) (video, 13:36). https://www.ted.com/talks/iyad_rahwan_what_moral_decisions_should_driverless_cars_make

This primer makes the case for a strict liability regime for autonomous vehicles as the best way to incentivize manufacturer to develop the safest possible self-driving cars. American Association for Justice, “Driven to Safety: Robot Cars and The Future of Liability,” (February 2017). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2913028

Labor and Employment

The use of artificial intelligence is widespread in recruiting and hiring. AI is being used to make individualized job suggestions, to screen resumes, and to assess candidates. Anyone who uses LinkedIn has received an algorithm-driven job recommendation, and with the recent launch of Google's Job Search engine we can expect AI to continue to shape the way companies recruit and hire.

The use of AI does not stop once a candidate is hired. Human resources departments are beginning to use AI in their evaluation and management of existing employees. A new AI tool on Slack analyzes messages between employees to alert managers when there is a change in morale. However, utilizing AI – whether in hiring or evaluating employees – raises issues of employee/candidate privacy and implicit bias.

The rapid spread of AI may exacerbate the problem of biased decision-making in recruiting and hiring practices. Another concern is that AI tools can access candidate information that would not typically be available to prospective employers. Dipayan Ghosh, "AI is the Future of Hiring, but It's Far From Immune to Bias," Quartz (October 17, 2017) <https://work.qz.com/1098954/ai-is-the-future-of-hiring-but-it-could-introduce-bias-if-were-not-careful/>

Human resources departments are beginning to use artificial intelligence tools to make decisions about hiring, firing and compensation. A company called Xander has developed a program that "can determine whether an employee feels, optimistic, confused or angry, and provide insights to help manage teams." But employment lawyers are concerned that use of these algorithms could reflect biases, and lead to employment discrimination. Imani Moise, "What's on Your Mind? Bosses Are Using Artificial Intelligence to Find Out," Wall St. Journal, (March 20, 2018). <https://www.wsj.com/articles/whats-on-your-mind-bosses-are-using-artificial-intelligence-to-find-out-1522251302>

A summary of a 2015 Carnegie Mellon study on how Google's algorithm recommended higher paying jobs to men than to women. Julia Carpenter, "Google's Algorithm Shows Prestigious Job Ads to Men, But Not to Women. Here's Why that Should Worry You," Washington Post (July 6, 2015) <https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/>

Privacy

The widespread adoption of Artificial Intelligence allows companies to not only gather more data on consumers, but also to draw more conclusions. Even as AI yields benefits for consumers, there is increased potential for breaches of consumer privacy. The readings below offer examples of the different contexts in which privacy concerns can arise:

AI is being used to detect individuals at risk for suicide. In one month late in 2017, Facebook's AI system, which analyzes posts and live streams alerted first responders in 100 cases. How will companies handle sensitive records about suicide risk and how can individuals go about erasing records of false positives? Aili McConnon, "AI Helps Identify People at Risk for Suicide," Wall St. Journal (February 23, 2018) <https://www.wsj.com/articles/ai-helps-identify-people-at-risk-for-suicide-1519400853>

"Deepfakes" are very realistic fake videos created by artificial intelligence. A program called FakeApp makes it relatively easy to superimpose one person's face on a video of another person. The main concern is that these deepfakes can seriously damage people's reputations, especially by creating fake revenge porn. Kevin Roose, "Here Come the Fake Videos, Too," NY Times (March 4, 2018) <https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html>

This academic article provides a brief overview of the ways in which AI has reshaped the consumer data privacy risks. Ginger Zhe Jin, "Artificial Intelligence and Consumer Protection," National Bureau of Economics (December 18, 2017). <http://www.nber.org/chapters/c14034.pdf>

Algorithmic Bias and Accountability

Because most AI depends in large part on identifying patterns based on data about past decisions, it risks applying and perpetuating longstanding biases. How can algorithms be held accountable when they make mistakes or lead to harmful consequences? What are the steps that can be taken during the design or implementation phase to increase algorithmic accountability?

This article discusses different degrees of algorithmic bias, ranging from unintentional bias to intentional and potentially illegal bias, and discusses ways to address them. Cathy O’Neil, “How can we stop algorithms telling lies?,” *The Guardian* (July 16, 2017) <https://www.theguardian.com/technology/2017/jul/16/how-can-we-stop-algorithms-telling-lies>

This article discusses some of the challenges in pattern identification and the resulting concerns about trust and explainability, focusing on image recognition as an example. Will Knight, “The Dark Secret at the Heart of AI.” *Technology Review*. (April 11, 2017) <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>

This piece lays out five main principles, and associated questions and suggested steps for those who design algorithms and those who use them in decision making. The principles are Responsibility, Explainability, Accuracy, Auditability and Fairness. Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H.V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Wu, Bendert Zevenbergen, “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms,” *Fairness, Accountability, and Transparency in Machine Learning* <https://www.fatml.org/resources/principles-for-accountable-algorithms>

This paper examines what standard of explanation should be required from algorithms. After reviewing the social, moral and legal norms around explaining human decision making, the paper argues that the same standard of explanation for human decision making should be applied to machine decision making. AI systems will therefore have to be designed to store inputs, intermediate steps and outputs in order to generate ex post explanations of their decisions. Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Stuart Schieber, James Waldo, David Weinberger, Alexandra Wood, “Accountability of AI Under the Law: The Role of Explanation,” *Arxiv.org*, [v2] (November 21, 2017) <https://arxiv.org/abs/1711.01134>

Government Use of Artificial Intelligence

As with the private sector, AI also has the potential to streamline and improve government decisionmaking and operations. Government applications for AI including providing automated constituent services, reviewing claims and complaints, allocating government services (e.g. benefits, street services, police patrols), and more. Government will need to consider and address many of the same concerns that face private users of AI (e.g. transparency, algorithmic bias), as well as unique public sector concerns (e.g. due process rights, democratic oversight). The readings below provide an overview of different ways governments might use AI:

This short article and accompanying infographic briefly identify 26 ways in which AI can help government. “Automation Beyond the Physical: AI in the Public.” Government Technology. (September 2017). <http://www.govtech.com/civic/GT-September-Automation-Beyond-the-Physical-AI-in-the-Public-Sector.html>

This joint report by IBM and the Partnership for Public Service provides an overview of how the federal government can utilize AI to improve its operations. “Using Artificial Intelligence to Transform Government.” The IBM Center for The Business of Government and the Partnership for Public Service. (January 2018). <http://www.businessofgovernment.org/sites/default/files/Using%20Artificial%20Intelligence%20to%20Transform%20Government.pdf>

This report by Deloitte discusses government applications of AI, including providing various frameworks for categorizing different roles for AI. “AI-augmented Government.” Deloitte Center on Government Insights. (2017) https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf

This academic paper by Prof. Mariano-Florentino Cuéllar (now a California Supreme Court Justice) explores trade-offs of delegating administrative agency decisions to AI. Mariano-Florentino Cuéllar. “Cyberdelegation and the Administrative State.” (January 2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2754385

This report by the AI NOW Institute at NYU proposes an “algorithmic impact assessment” as a way of improving how governments plan for, adopt, and use algorithmic decision-making systems, including AI technologies. Dillon Reisman, Jason Schultz, Kate Crawford, & Meredith Whittaker. “Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability.” (April 2018). <https://ainowinstitute.org/aiareport2018.pdf>

Antitrust and Market Fairness

AI has the potential to implicate antitrust concerns, a longstanding area of AG jurisdiction. Because AI generally improves the larger the amount of available data it has from which to learn patterns, the growing importance of AI to the economy may entrench larger and/or incumbent market actors:

A longer-term concern is the way AI creates a virtuous circle or “flywheel” effect, allowing companies that embrace it to operate more efficiently, generate more data, improve their services, attract more customers and offer lower prices. That sounds like a good thing, but it could also lead to more corporate concentration and monopoly power—as has already happened in the technology sector.¹

AI will also allow companies to better coordinate their activities—often in ways that evade detections. More immediately, companies have already begun to engage in “tacit collusion,” using algorithms to constantly adjust their prices. As these algorithms become more sophisticated, regulators may have to shift their focus from policing their activities to mandating “compliance by design”.

For more on the implications of AI for antitrust enforcement, please see the resources below:

Michaela Ross, “Artificial Intelligence Pushes the Antitrust Envelope.” Bloomberg Law. (April 28, 2017) <https://www.bna.com/artificial-intelligence-pushes-n57982087335/>

Nicholas Hirst, “When Margrethe Vestager takes antitrust battle to robots.” POLITICO. (February 18, 2018) <https://www.politico.eu/article/trust-busting-in-the-age-of-ai/>

This ITIF paper challenges common concerns about the antitrust implications of data collection and AI. Joe Kennedy, “The Myth of Data Monopoly: Why Antitrust Concerns About Data Are Overblown.” (March 2017). <http://www2.itif.org/2017-data-competition.pdf>

This filing by the FTC briefly discusses algorithms and relevant antitrust case law. “Algorithms and Collusion - Note by the United States.” Organisation for Economic Co-operation and Development. (26 May 2017). [6 pages] <https://www.ftc.gov/system/files/attachments/us-submissions-oecd-other-international-competition-fora/algorithms.pdf>

¹ <https://www.economist.com/news/special-report/21739431-artificial-intelligence-spreading-beyond-technology-sector-big-consequences>

Broader AI Effects and Proposed Regulation

AI will have far-reaching effects beyond the scope of traditional AG jurisdiction. Yet these impacts—including for example job displacement, the restructuring of transportation infrastructure, and the continuing evolution of social media and online communities—will shape the underlying political landscape. These changes may have indirect effects on areas of AG concern, and affected constituencies may generate pressure for AGs to apply additional scrutiny to actors who deploy AI.

The readings below provide an overview of some of the impact AI is expected to have on our society and economy.

This report by the Obama Administration provides a broad policy agenda for maximizing the positive effects of AI and ensuring it benefits the broadest number of people. The White House. Artificial Intelligence, Automation, and the Economy. (December 2016) [55 pages] <https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF>

This report takes an optimistic view, challenging many of the most common fears about the negative impact of artificial intelligence. Robert D. Atkinson, “It’s Going to Kill Us!’ and Other Myths About the Future of Artificial Intelligence,” Information Technology & Innovation Foundation (June 2016) <http://www2.itif.org/2016-myths-machine-learning.pdf>

This article walks through major policy challenges and ethical questions raised by AI generally. Ryan Calo, “Artificial Intelligence Policy: A Primer and Roadmap,” (October 17, 2017) [28 pages] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015350

Presenter Materials: Assembly's Data Nutrition Label

■ on the web: datanutrition.media.mit.edu

The Dataset Nutrition Label:

A Framework to Drive Higher Data Quality Standards

Sarah Holland^{1*}, Ahmed Hosny^{2*}, Sarah Newman³, Joshua Joseph⁴, and Kasia Chmielinski¹
April 2018 | nutrition@media.mit.edu | datanutrition.media.mit.edu

¹Assembly, MIT Media Lab and Berkman Klein Center for Internet & Society at Harvard University, ²Dana-Farber Cancer Institute, Harvard Medical School, ³metaLAB (at) Harvard, Berkman Klein Center for Internet & Society, Harvard University, ⁴33x.ai

*Authors contributed equally

ABSTRACT

Artificial intelligence (AI) systems matter, and so does the data on which they are modeled. One way to improve the accuracy and fairness of these models -- models that determine everything from navigation directions to mortgage approvals -- is to make it easier for data scientists to quickly assess the viability and fitness of datasets used to train them. Current methods of data analysis and assessment are not standardized; they vary greatly across industry and domains, and are also costly in time and expertise. Drawing from the fields of nutrition and privacy, we introduce a solution: the Dataset Nutrition Label, a diagnostic framework to address and mitigate some of the challenges in this process.

The Dataset Nutrition Label (the Label) provides data scientists with a distilled yet comprehensive overview of dataset ‘ingredients’ before AI model development. The Label is comprised of modules that indicate key dataset attributes in a standardized format. The Label also utilizes multiple statistical and probabilistic modelling backends to capture such attributes. The modules include both qualitative and quantitative information, and vary in both the access they require to the dataset as well as the point at which they are generated or appended. Different combinations of modules can be used to customize Labels for particular datasets. Modules can also be added at a later time by data scientists using and interrogating a dataset. To demonstrate and advance this concept, we generated a Label for the ProPublica *Dollars for Docs* dataset, which documents payments made to physicians by drug companies in the U.S. between 2013-2015. The live, interactive, and open source prototype displays the following sample modules: metadata, data provenance, diagnostic statistics, variable correlations, and ground truth comparisons. The prototype is available on the Dataset Nutritional Label website.

The Dataset Nutrition Label offers many benefits. It drives robust data analysis practices by providing a pre-generated ‘floor’ for basic data interrogation. Data scientists selecting datasets for model development can leverage the Label to quickly compare the ‘health’ of multiple datasets, helping them efficiently select the best dataset for their purposes, and avoiding onerous and costly analyses. Improved dataset selection provides a secondary benefit: the quality of the models trained on that data will also improve. This is a result of using a more robust dataset generally, and also as the Label enables data scientists to check for additional issues at the time

of model development (e.g. surprising variable correlations, missing data, anomalous data distributions, etc). The existence of these Labels will also prompt data scientists to question a dataset and its characteristics regardless of whether every dataset contains such a Label. Lastly, by encouraging the authors and users of datasets to create Labels, we hope to build awareness of significant shortcomings in datasets (e.g. missing data, biased collection practices), which in turn will drive better data collection practices and more responsible dataset selection and use in the future.

We also explore the limitations of the Label. Considering the variety of datasets used to build models today, some challenges arise with generalizing the Label across data type, size, and composition. Some of the modules require access to the dataset's author and the data itself; this could constrain the creation of certain modules to only those who own, manage, or have access to the data. We discuss the risk of modules that rely on 'ground truth' data, which will depend on the accuracy of such ground truth for comparison. The design of the Label itself will require additional attention to determine the appropriate amount of information for presentation, comprehension, and adoption. Finally, we discuss potential ways to move forward given the limitations identified.

The Dataset Nutrition Label is a useful, timely, and necessary intervention in the development of AI models: it will encourage the collection of better and more complete data and more responsible usage of such data; it will drive accountability across various industries, and it will mitigate harm caused by algorithms built on problematic or ill-fitting data. We lay out future directions for the Dataset Nutrition Label project, including research and public policy agendas to further advance consideration of dataset labeling.

KEYWORDS

Dataset, data, quality, bias, analysis, artificial intelligence, machine learning, model development, nutrition label, computer science, data science, governance, AI accountability, algorithms

THE DATASET NUTRITION LABEL PROJECT

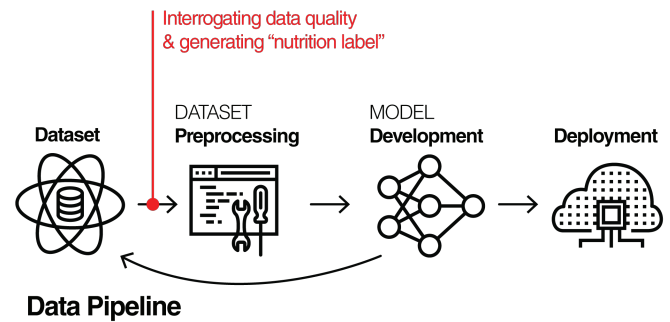
Empowering data scientists and policymakers with practical tools to improve AI outcomes

datanutrition.media.mit.edu | nutrition@media.mit.edu

The Problem - Garbage in, Garbage out

Algorithms matter, and so does the data they're trained on. To improve the accuracy and fairness of algorithms that determine everything from navigation directions to mortgage approvals, we need to make it easier for practitioners to quickly assess the viability and fitness of datasets they intend to train AI algorithms on.

There's a missing step in the AI development pipeline: assessing datasets based on standard quality measures that are both qualitative and quantitative. We are working on packaging up these measures into an easy to use Dataset Nutrition Label.

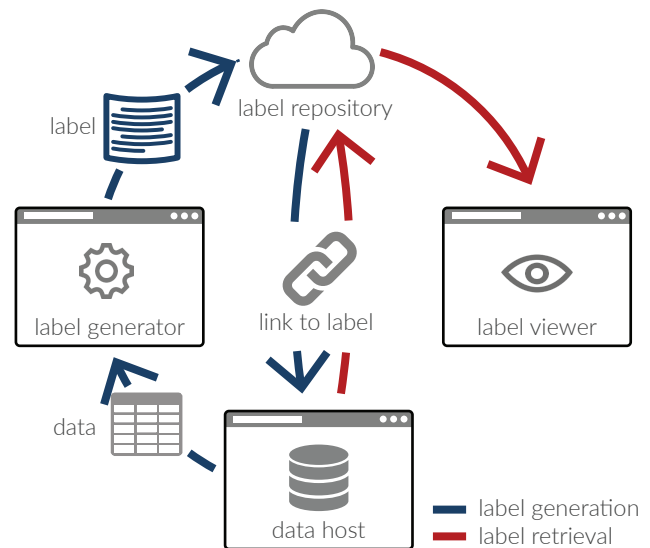


Our Solution - A Nutrition Label for Datasets

The Dataset Nutrition Label aims to create a standard for interrogating datasets for measures that will ultimately drive the creation of better, more inclusive algorithms.

Our current prototype includes a highly-generalizable interactive data diagnostic label that allows for exploring any number of domain-specific aspects in datasets. Similar to a nutrition label on food, our label aims to highlight the key ingredients in a dataset such as meta-data and populations, as well as unique or anomalous features regarding distributions, missing data, and comparisons to other 'ground truth' datasets. We are currently testing our label on several datasets, with an eye towards open sourcing this effort and gathering community feedback.

The design utilizes a 'modular' framework that can be leveraged to add or remove areas of investigation based on the domain of the dataset. For example, Nutrition Labels for data about people may include modules about the representation of race and gender, while Nutrition Labels for data about trees may not require that module.



THE DATASET NUTRITION LABEL PROJECT

Empowering data scientists and policymakers with practical tools to improve AI outcomes

datanutrition.media.mit.edu | nutrition@media.mit.edu

Label Prototype

We developed this prototype on ProPublica's Dollars for Docs (2013-2015) dataset, which details payments made from pharmaceutical companies to doctors. It includes several 'modules' across a variety of qualitative and quantitative data that we believe is useful for exploring several aspects in datasets before the development of models.

Dataset Facts

ProPublica's Dollars for Docs Data

Metadata

Filename	201612v1-dolldollars-product_payments
Format	csv
Uri	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015

Description This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

Provenance

Source	
Name	U.S. Centers for Medicare & Medicaid Services
Uri	https://www.cms.gov/OpenPayments/
Email	openpayments@cms.hhs.gov
Author	
Name	Propublica
Uri	https://www.propublica.org/dataset/
Email	data.store@propublica.org

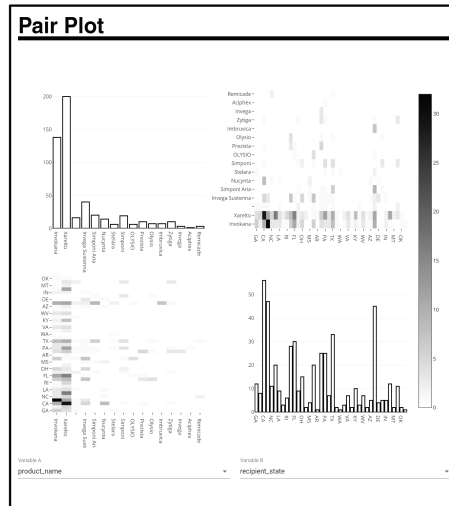
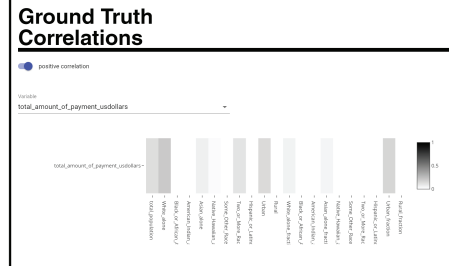
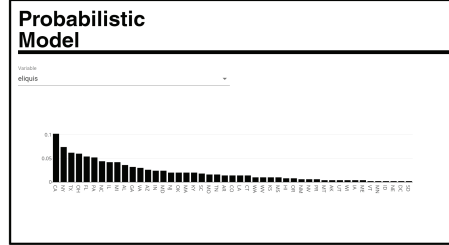
Statistics

Ordinal							
name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing	
id	number	500	499 including mis...	missing value (13)	multiple detected	2.6%	
applicable_man...	number	500	4	10000000032 (multiple detected	0%	
date_of_payment	date	500	213 including mis...	missing value (27)	multiple detected	5.4%	
general_transac...	number	500	487 including mis...	missing value (34)	multiple detected	6.8%	
program_year	number	500	2 including missi...	2014 (495)	missing value (5)	1.0%	

Nominal							
name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing	
product_name	string	500	16 including mis...	Xarelto (200)	Aspich (1)	3.2%	
original_product...	string	500	15	Xarelto (212)	Aspich (1)	0%	
product_ndc	number	500	21 including miss...	5043857810 (207)	multiple detected	5.0%	
product_is_drug	boolean	500	2 including miss...	1 (492)	missing value (9)	1.8%	
payment_has_m...	boolean	500	3 including miss...	1 (397)	missing value (99)	5.8%	
teaching_hospit...	number	500	2 including miss...	0 (444)	missing value (36)	7.2%	
physician_profil...	number	500	230 including mis...	missing value (92)	multiple detected	6.4%	
recipient_state	string	500	40	CA (55)	multiple detected	0%	
applicable_man...	string	500	5 including miss...	Janssen Pharm...	multiple detected	7.0%	
teaching_hospit...	number	500	2 including miss...	0 (481)	missing value (19)	3.8%	
product_slug	string	500	15 including mis...	drugxarelto (196)	drug-aspich (1)	8.2%	

Continuous									
name	type	count	min	median	max	mean	standardD...	missing	zeros
total_amo...	number	500	0.14	14.00	5000	134.21	501.99	9.40%	0%

Discrete									
name	type	count	min	median	max	mean	standardD...	missing	zeros
number_o...	number	500	1.00	1	1	1.00	0.00	4.80%	0%



Variables

id	A unique ID number for this payment & product combination. This is assigned by ProPublica for internal use
Applicable_manufacturer_or_applicable_gpo_making_payment_id	ID of the applicable manufacturer or submitting applicable GPO making the payment or other transfer of value
Date_of_payment	If a singular payment, then this is the actual date the payment was issued; if a series of payments or an aggregated set of payments, this is the date of the first payment to the covered recipient in this program year
General_transaction_id	System-assigned identifier to the general transaction at the time of submission
Program_year	The calendar year for which the payment is reported in Open Payments
Product_name	Derived from the 'name_of_associated_covered_drug_or_biologicalX' field (for drugs) or 'name_of_associated_covered_device_or_medical_supplyX' field (for medical devices). Where possible, multiple versions of the same product are converted to the same product_name (i.e. records for 'Zorvolex 65mg' and 'Zorvolex 35mg' will be converted to 'Zorvolex'). The original value is contained in original_product_name
Original_product_name	A copy of the original 'name_of_associated_covered_drug_or_biologicalX' field (for drugs) or 'name_of_associated_covered_device_or_medical_supplyX' field (for medical devices)
Product_ndc	If the product is a drug, this a copy of the original 'ndc_of_associated_covered_drug_or_biologicalX' field
Product_is_drug	't' if the product is a drug (contained in a 'name_of_associated_covered_drug_or_biologicalX' field). 'f' if the product is a medical device (contained in a 'name_of_associated_covered_device_or_medical_supplyX' field)
Payment_has_many	't' if the original payment record included data on more than one drug or device, i.e. 'name_of_associated_covered_drug_or_biologicalal1' and 'name_of_associated_covered_drug_or_biologicalal2', 'name_of_associated_covered_device_or_medical_supply1' and 'name_of_associated_covered_device_or_medical_supply2', etc.
Teaching_hospital_id	Open Payments system-generated unique identifier of the teaching hospital receiving the payment or other transfer of value
Physician_profile_id	ID of the physician receiving the payment or other transfer of value
Recipient_state	The state or territory abbreviation of the primary business address of the physician or teaching hospital or non-covered recipient entity receiving the payment or other transfer of value if the primary business address is in the United States
Applicable_manufacturer_or_applicable_gpo_making_payment_name	Textual proper name of the applicable manufacturer or applicable GPO making the payment or other transfer of value. This field has been standardized to eliminate different names attributable solely to punctuation
Teaching_hospital_ccn	A unique identifying number (CMS Certification Number) of the Teaching Hospital receiving the payment or other transfer of value
Product_slug	Used internally at ProPublica for web display on the Dollars for Docs app. You can pull up the corresponding Dollars for Docs page for a product by appending product_slug to https://projects.propublica.org/docdollars/products/ , i.e. https://projects.propublica.org/docdollars/products/device-cental-cabinetry
Total_amount_of_payment_usdollars	U.S. dollar amount of payment or other transfer of value to recipient (manufacturer must convert to dollar currency if necessary)
Number_of_payments_included_in_total_amount	The number of discrete payments being reported in the 'Total Amount of Payment' data element

