Bloomberg Law Tech & Telecom

Blockchain 'Smart Contracts' – A New Transactional Framework

*By **Laura E. Jehl** and **Brian Bartish***
Baker & Hostetler LLP

This article, the third in a series, explores applications of blockchain technology.

With the growing buzz around blockchain technology, many organizations are in a race to position themselves as early adopters and leaders in the space. For these organizations, one of the more exciting blockchain applications is the promise of increased efficiency and reduced costs in the transacting process through so-called "smart contracts" – which are actually neither "smart" nor necessarily true legal contracts. Smart contracts are automated programs that encode transactional logic for self-execution and rely upon decentralized cryptographic methods to effectuate enforcement. Regardless of one's opinion of their name, or their legal status, smart contracts are garnering a significant amount of attention and investment due to their ability to radically transform the way parties transact with one another.

♦ What are smart contracts?

Smart contracts actually predate the creation of blockchain technology, as the term "smart contracts" was first coined by computer science and legal researcher Nick Szabo in the mid-1990s. Szabo defined a smart contract as "a set of promises, specified in digital form, including protocols within which the parties perform on these promises." He offered an analogy of the vending machine to illustrate his premise that the entire environment of the transaction could be created within the purview of a machine. In stocking the vending machine, the owner has created an offer, which is accepted when a buyer inserts cash and makes a selection. The code running the machine then takes over to perfect performance by verifying the currency input, dispensing the buyer's selection, and returning any required change. While Szabo's initial vision of smart contracts promised modest gains in transactional efficiency through automation, the advent of blockchain technology has created a number of significant new benefits, perhaps chief among them the ability to create trust between parties operating in a trustless environment that does not rely on a centralized institution, government, or other middleman.

♦ How do smart contracts operate?

Smart contracts rely on code deployed on a blockchain to automatically execute the terms of an agreement. This is where smart contracts begin to depart from traditional contracts, i.e., agreements embodying certain terms to be fulfilled by parties and given the force of law to incentivize performance. In contrast, smart contracts can be viewed as "autonomous agents" designed to execute the logic of an agreement through code that responds to specific messages or transactions. In computational terms, smart contracts are programs that can execute an arbitrary, or open-ended, array of user-specified state transition functions, including performing calculations and storing information. These alter the collective status, or state, of the underlying system, which embodies the entire history of preceding events and the way those events bear upon circumstances such as the ownership of outstanding currencies, the location of goods in transit, or the status of voting rights. The programs function as cryptographic "boxes" that

contain value or information and that can only be unlocked in response to certain predefined conditions. Smart contracts, therefore, aren't truly smart, but rather deterministic.

While not as "smart" as advertised, blockchain-based smart contracts represent an evolution of the underlying bitcoin technology, requiring more powerful platforms and more robust programming languages. The Ethereum Virtual Machine, or simply Ethereum, is the best-known of these platforms. Ethereum emerged with its own programming language, Solidity, specifically designed to encode logic into smart contracts. Ethereum and Solidity offer important advancements over the bitcoin architecture, as both were designed to be "Turing-complete," meaning that they can encode any computation that can be conceivably carried out, including infinite loops. This capability becomes important as the complexity of smart contracts increases, particularly when a smart contract calls on another smart contract as an independent data source or a verifier of real-world events (often referred to as an "oracle"). For example, smart contracts involving financial derivatives may rely on an external source of data, such as the value of the dollar or the Nasdaq index, which can be fed to the derivatives contract through a separate smart contract deployed specifically for calculating those functions. The fact that Solidity is Turing-complete, however, may expose users to infinite loops in contract execution that can cause significant delays and waste both computational and financial resources. Ethereum attempts to manage this type of "denial of service" threat through its transaction structure. Each Ethereum transaction consists of:

- the message recipient;
- the cryptographic signature of the sender;
- an amount of ether (the cryptocurrency used on Ethereum) to transfer;
- an optional data field;
- a "startgas" value, which represents the maximum number of computational steps that a transaction can take when executing; and
- a "gasprice" value, which represents the price per computational step that the sender pays to the miner in order to publish the transaction to the blockchain.

In the event that a transaction "runs out of gas" before completing its execution, the participating nodes and the entire blockchain revert to their previous states, but the miner (i.e., the node that earns the right to publish the block containing the transaction) still collects the gasprice transaction fee. This design, however, is not foolproof against all malicious attacks and still presents some significant risks due, in part, to simple programming error.

Another risk was exemplified by the so-called Decentralized Autonomous Organization (DAO), where a number of Ethereum users joined together to create a sort of crowd-funded venture capital fund where members could vote to invest the DAO's funds in a number of projects. This early attempt at an organization managed entirely through smart contracts ended in ignominy, however, as an attacker exploited flaws in the logic of the underlying smart contracts to siphon off nearly $50 million in ether. The funds were recovered, but only after Ethereum leaders convinced a majority of nodes on the platform to implement a "hard fork" – essentially an operation that reverted the state of the network to what it was prior to the theft. This hard fork, however, required the users to abandon the original network, which still exists under the name Ethereum Classic. The DAO hack served as a lesson that many of the purported strengths of the blockchain architecture, such as its immutability, may be detrimental in certain contexts. Users should therefore carefully consider whether the blockchain will increase the efficiency of transactions or subject them to heightened or unnecessary risks.

♦ Smart contract use cases

Despite the risks, smart contracts offer a number of exciting potential use cases. The developers of Ethereum envisioned a broad array of uses, such as financial derivatives for crop insurance, savings wallets, wills, employment contracts, and peer-to-peer gambling. Smart contract use cases extend beyond the purely financial, as they offer a potential solution to coordination failures among transacting parties. They also offer avenues for experimentation with decentralized governance structures for software development, project management, and entire business organizations.

Unlike the early days of Ethereum, corporations are now investing in smart contract pilots and setting up joint ventures to work on the technology. In 2017, AIG partnered with IBM to create a smart contract multinational insurance policy for Standard Chartered Bank PLC. The policy operates through multiple

smart contracts, covering a main policy for Standard Chartered's U.K. headquarters and local policies for affiliates in the U.S., Singapore, and Kenya, which communicate to share data and documents. Also in 2017, French insurer AXA started testing Fizzy, a flight-delay insurance product that leverages smart contracts on the Ethereum blockchain. The smart contracts are connected to global air traffic databases so that as soon as a flight is delayed more than two hours, the smart contract triggers compensation to the insured traveler.

One of the most oft-cited implementations of a smart contract is supply chain management, in which a contract or series of contracts is part of a system that automatically controls the shipment of goods and payments through all stages of the logistics cycle. IBM recently announced a new joint venture with Danish firm A.P. Moller-Maersk – the world's largest container shipping firm, handling roughly one in seven containers shipped globally – that will implement smart contracts as part of a comprehensive strategy to digitize the global supply chain. Their goal is to drive down expenses and increase the speed of the end-to-end shipping process by using smart contracts to automate costly customs clearance and approval requirements.

Beyond the corporate world, governments are also experimenting with the technology. Sweden's land registry authority, the Lantmäteriet, is testing a system for real estate transactions and mortgage deed processes. This would allow buyers and sellers to strike a deal using a smart contract connected to a private blockchain, which reduces the need for paperwork and provides greater transparency in chain of title. One of the hurdles in the Lantmäteriet's road map is a legal issue: validity of digital signatures for real estate contracts. Elsewhere around the globe, Dubai is undertaking a comprehensive digital transformation that would migrate all visa applications, bill payments, and license renewals to blockchain technology by 2020.

While blockchain-based smart contracts are still in a state of infancy and their risks are not always fully anticipated, the interest in their applications to the commercial sector has intensified development efforts. Hyperledger, Project Accord, and the Enterprise Ethereum Alliance have already gained a number of influential supporters from various fields.

Hyperledger is a membership-based organization with the objective of advancing cross-industry blockchain technologies. It incubates and promotes a number of tools including Hyperledger Burrow – a smart contract machine contributed by smart contract startup Monax and co-sponsored by Intel – which executes Ethereum smart contract code on a permissioned virtual machine. Hyperledger has more than 100 members, from tech companies to banks and academic institutions to commercial industry groups.

The Accord Project is an open source software initiative established with Hyperledger, the International Association for Contract & Commercial Management, and the W3C, a web standards body. One of its projects, Cicero, aims to provide lawyers and business professionals with a system for turning paper-based, legally binding agreements into legally binding smart contracts. The Accord Project's membership consists of big law firms, startups, venture capital firms, and other organizations.

More than 150 organizations from a range of industries – including software, infrastructure, financial services, manufacturing, and law – signed on to the Enterprise Ethereum Alliance, launched in February 2017. Formed with the goal of connecting business leaders, startups, academics, and vendors with Ethereum subject matter experts to establish a road map for enterprise adoption, the Ethereum Enterprise Alliance counts Microsoft, JPMorgan Chase, Mastercard, BP, ING, and Deloitte among its members.

Propelled by the strong interest of these well-funded industry leaders, smart contracts are increasingly appearing on legislative agendas. Arizona and Nevada have recently passed laws that promote the legal enforceability of smart contracts, and Florida appears poised to do the same. Much like smart contracts themselves, the gears of progress propelling these efforts appear poised to self-execute.

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 Comande [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 on 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 important 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

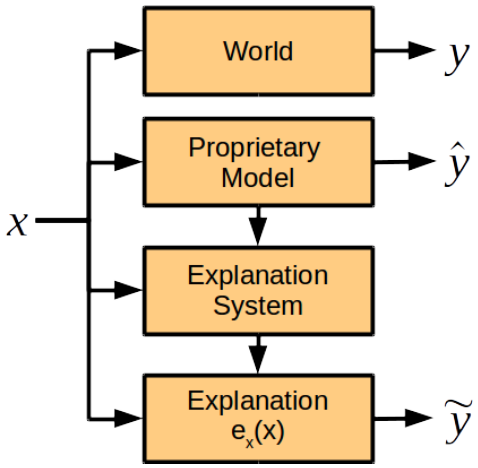

Figure 1: Diagram of a Framework for Explainable AI Systems.

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

	Human	*AI*
Strengths	Can provide explanation post-hoc	Reproducible, no social pressure
Weaknesses	May be inaccurate and irreliable, 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

Approach	*Well-suited Contexts*	*Poorly-suited Contexts*
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 (ICRCICN), 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 Comande. 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 Prodcut 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. Patternnet 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.

BROOKINGS

TechTank

Artificial intelligence and bias: Four key challenges

John Villasenor Thursday, January 3, 2019

It is not news that, for all its promised benefits, artificial intelligence has a bias problem. Concerns regarding racial or gender bias in AI have arisen in applications as varied as hiring, policing, judicial sentencing, and financial services. If this extraordinary technology is going to reach its full potential, addressing bias will need to be a top priority. With that in mind, here are four key challenges that AI developers, users, and policymakers can keep in mind as we work to create a healthy AI ecosystem.

1. Bias built into data

We live in a world awash in data. In theory, that should be a good thing for AI: After all, data give AI sustenance, including its ability to learn at rates far faster than humans. However, the data that AI systems use as input can have built-in biases, despite the best efforts of AI programmers.

Consider an algorithm used by judges in making sentencing decisions. It would obviously be improper to use race as one of the inputs to the algorithm. But what about a seemingly race-neutral input such as the number of prior arrests? Unfortunately, arrests are not race neutral: There is plenty of evidence indicating that African-Americans are disproportionally targeted in policing. As a result, arrest record statistics are heavily shaped by race. That correlation could propagate in sentencing recommendations made by an AI system that uses prior arrests as an input.

The indirect influence of bias is present in plenty of other types of data as well. For instance, evaluations of creditworthiness are determined by factors including employment history and prior access to credit—two areas in which race has a major impact. To take another example, imagine how AI might be used to help a large company set starting

salaries for new hires. One of the inputs would certainly be salary history, but given the well-documented concerns regarding the role of sexism in corporate compensation structures, that could import gender bias into the calculations.

2. AI-induced bias

An additional challenge is that biases can be created within AI systems and then become amplified as the algorithms evolve.

By definition, AI algorithms are not static. Rather they learn and change over time. Initially, an algorithm might make decisions using only a relatively simple set of calculations based on a small number of data sources. As the system gains experience, it can broaden the amount and variety of data it uses as input, and subject those data to increasingly sophisticated processing. This means that an algorithm can end up being much more complex than when it was initially deployed. Notably, these changes are not due to human intervention to modify the code, but rather to automatic modifications made by the machine to its own behavior. In some cases, this evolution can introduce bias.

Take as an example software for making mortgage approval decisions that uses input data from two nearby neighborhoods—one middle-income, and the other lower-income. All else being equal, a randomly selected person from the middle-income neighborhood will likely have a higher income and therefore a higher borrowing capacity than a randomly selected person from the lower-income neighborhood.

Now consider what happens when this algorithm, which will grow in complexity with the passage of time, makes thousands of mortgage decisions over a period of years during which the real estate market is rising. Loan approvals will favor the residents of the middle-income neighborhood over those in the lower-income neighborhood. Those approvals, in turn, will widen the wealth disparity between the neighborhoods, since loan recipients will disproportionally benefit from rising home values, and therefore see their future borrowing power rise even more.

Analogous phenomena have long occurred in non-AI contexts. But with AI, things are far more opaque, as the algorithm can quickly evolve to the point where even an expert can have trouble understanding what it is actually doing. This would make it hard to know if it is engaging in an unlawful practice such as redlining.

The power of AI to invent algorithms far more complex than humans could create is one of its greatest assets—and, when it comes to identifying and addressing the sources and consequences of algorithmically generated bias, one of its greatest challenges.

3. Teaching AI human rules

From the standpoint of machines, humans have some complex rules about when it is okay to consider attributes that are often associated with bias. Take gender: We would rightly deem it offensive (and unlawful) for a company to adopt an AI-generated compensation plan with one pay scale for men and a different, lower pay scale for women.

But what about auto insurance? We consider it perfectly normal (and lawful) for insurance companies to treat men and women differently, with one set of rates for male drivers and a different set of rates for female drivers—a disparate treatment that is justified based on statistical differences in accident rates. So does that mean it would be acceptable for an algorithm to compute auto insurance rates based in part on statistical inferences tied to an attribute such as a driver's religion? Obviously not. But to an AI algorithm designed to slice enormous amounts of data in every way possible, that prohibition might not be so obvious.

Another example is age. An algorithm might be forgiven for not being able to figure out on its own that it is perfectly acceptable to consider age in some contexts (e.g., life insurance, auto insurance) yet unlawful to do so in others (e.g., hiring, mortgage lending).

The foregoing examples could be at least partially mitigated by imposing upfront, application-specific constraints on the algorithm. But AI algorithms trained in part using data in one context can later be migrated to a different context with different rules about

the types of attributes that can be considered. In the highly complex AI systems of the future, we may not even know when these migrations occur, making it difficult to know when algorithms may have crossed legal or ethical lines.

4. Evaluating cases of suspected AI bias

There is no question that bias is a significant problem in AI. However, just because algorithmic bias is suspected does not mean it will actually prove to be present in every case. There will often be more information on AI-driven outcomes—e.g., whether a loan application was approved or denied; whether a person applying for a job was hired or not—than on the underlying data and algorithmic processes that led to those outcomes. This can make it more difficult to distinguish apparent from actual bias, at least initially.

While an accusation of AI bias should always be taken seriously, the accusation itself should not be the end of the story. Investigations of AI bias will need to be structured in a way that maximizes the ability to perform an objective analysis, free from pressures to arrive at any preordained conclusions.

The upshot

While AI has the potential to bring enormous benefits, the challenges discussed above—including understanding when and in what form bias can impact the data and algorithms used in AI systems—will need attention. These challenges are not a reason to stop investing in AI or to burden AI creators with hastily-drafted, innovation-stifling new regulations. But they do mean that it will be important to put real effort into approaches that can minimize the probability that bias will be introduced into AI algorithms, either through externally-supplied data or from within. And, we'll need to articulate frameworks for assessing whether AI bias is actually present in cases where it is suspected.

BROOKINGS

TechTank

Artificial intelligence and the future of geopolitics

John Villasenor Wednesday, November 14, 2018

On September 1, 2017, Russian President Vladimir Putin addressed a nationwide group of Russian students on their first day of school. "Artificial intelligence is the future, not only for Russia, but for all humankind," he said. "Whoever becomes the leader in this sphere will become the ruler of the world."

It is an assessment that would be unwise to ignore, even if the goal is simply economic prosperity as opposed to world domination. Over the coming decade, AI will become linked with geopolitics to a level that is difficult to fully comprehend today. Why? Because geopolitics is determined in large part by many of the same domains that AI is poised to revolutionize.

AI will make manufacturing, transportation, and trade more efficient, improve crop yields, open a wealth of new opportunities for technology advances, reshuffle labor markets, and force a fundamental rethinking of approaches to national security and the architecture of modern militaries. In the coming decades, countries that are able to successfully cultivate and harness a culture of AI innovation will be well positioned for both economic growth and improved national security. By contrast, countries that maintain an overreliance on legacy infrastructure and economic models will face increasing challenges in sustaining global competitiveness.

U.S. leadership in AI

It is also important to emphasize that, geopolitically speaking, AI is not a zero-sum game. Much ink has been spilled recently over China's extraordinary level of investment in AI, often accompanied by the implication that Chinese AI progress will inevitably come at the expense of the United States. But that logic implies a non-existent causality. China is

betting on AI because its political and business leaders have correctly identified it as a critical element of continued Chinese economic growth. That in no way inhibits the United States from making its own investments in AI.

And, that is exactly what is occurring. The United States is a global AI leader, with an ecosystem that includes not only extensive AI research at major companies like Google, Amazon, Facebook, Apple, and IBM, but also hundreds of AI-focused startups in areas ranging from drug discovery to education to manufacturing. The collective American commercial sector investment going into AI is immense. The U.S. government is investing as well. In September 2018, DARPA announced a "$2 Billion Campaign to Develop Next Wave of AI Technologies" with the aim of "transforming computers from specialized tools to partners in problem-solving."

Furthermore, it's not only about dollars. The United States leads the world in AI human capital—an advantage that stands to grow even further given the extensive efforts in U.S. universities to ramp up research and teaching in AI and related topics.

In fact, some of the biggest potential AI challenges in the United States are actually at the level of policy and not technology or human capital. Maintaining AI preeminence is a multi-decade endeavor—a far greater time scale than the term lengths of elected officials. This lowers the incentives to implement AI-focused policy strategies that might take several years or more to bear fruit. Overregulation is another threat to American AI innovation, as it could hamper both the incentives to develop new AI technologies as well as the speed of delivering them to the marketplace.

A global AI ecosystem

While the United States and China are the largest AI players, the ecosystem is global. Israel and the United Kingdom have thriving AI sectors. Earlier this year, the French government announced a major public investment in AI. Facilitating AI innovation is also a key focus of governments in Japan, South Korea, and Russia.

Many of the benefits of country-specific AI investments will be both national and global. AI will make it easier to predict violent storms. It can help with drug development to help reduce the impact of disease. It can improve agricultural yields, and help manage the complexities of the supply chain for food, medicine, and other goods. All of these things have profoundly important geopolitical implications. In a September 2018 *Washington Post* piece, Nicolas Berggruen and Nathan Gardels wrote that "artificial intelligence has become the most powerful resource that will determine the fate of nations in the times ahead." It's an astute observation that will likely prove true.

AI is not magic, and there is plenty it cannot do. But AI is perhaps the only technology in recent memory that, despite all the hype, will actually turn out to have been underhyped once its impacts are fully appreciated. And while the full future impact of AI is impossible to predict, one thing is clear: As we move towards the middle of the 21st century, a nation's geopolitical standing and its strength in AI will be increasingly intertwined. It's a correlation that leaders across the globe will surely have in mind as they work to achieve their geopolitical aspirations.

BROOKINGS

TechTank

Artificial intelligence, deepfakes, and the uncertain future of truth

John Villasenor Thursday, February 14, 2019

Deepfakes are videos that have been constructed to make a person appear to say or do something that they never said or did. With artificial intelligence-based methods for creating deepfakes becoming increasingly sophisticated and accessible, deepfakes are raising a set of challenging policy, technology, and legal issues.

Deepfakes can be used in ways that are highly disturbing. Candidates in a political campaign can be targeted by manipulated videos in which they appear to say things that could harm their chances for election. Deepfakes are also being used to place people in pornographic videos that they in fact had no part in filming.

Because they are so realistic, deepfakes can scramble our understanding of truth in multiple ways. By exploiting our inclination to trust the reliability of evidence that we see with our own eyes, they can turn fiction into apparent fact. And, as we become more attuned to the existence of deepfakes, there is also a subsequent, corollary effect: they undermine our trust in *all* videos, including those that are genuine. Truth itself becomes elusive, because we can no longer be sure of what is real and what is not.

What can be done? There's no perfect solution, but there are at least three avenues that can be used to address deepfakes: technology, legal remedies, and improved public awareness.

While AI can be used to make deepfakes, it can also be used to detect them. Creating a deepfake involves manipulation of video data—a process that leaves telltale signs that might not be discernable to a human viewer but that sufficiently sophisticated detection algorithms can aim to identify.

As research led by professor Siwei Lyu of the University at Albany has shown, face-swapping (editing one person's face onto another person's head) creates resolution inconsistencies in the composite image that can be identified using deep learning techniques. Professor Edward Delp and his colleagues at Purdue University are using neural networks to detect the inconsistencies across the multiple frames in a video sequence that often result from face-swapping. A team including researchers from UC Riverside and UC Santa Barbara has developed methods to detect "digital manipulations such as scaling, rotation or splicing" that are commonly employed in deepfakes.

The number of researchers focusing on deepfake detection has been growing, thanks in significant part to DARPA's Media Forensics program, which is supporting the development of "technologies for the automated assessment of the integrity of an image or video." However, regardless of how far technological approaches for combating deepfakes advance, challenges will remain.

Deepfake detection techniques will never be perfect. As a result, in the deepfakes arms race, even the best detection methods will often lag behind the most advanced creation methods. Another challenge is that technological solutions will have no impact when they aren't used. Given the distributed nature of the contemporary ecosystem for sharing content on the internet, some deepfakes will inevitably reach their intended audience without going through detection software.

More fundamentally, will people be more likely to believe a deepfake or a detection algorithm that flags the video as fabricated? And what should people believe when different detection algorithms—or different people—render conflicting verdicts regarding whether a video is genuine?

The legal landscape related to deepfakes is complex. Frameworks that can potentially be asserted to combat deepfakes include copyright, the right of publicity, section 43(a) of the Lanham Act, and the torts of defamation, false light, and intentional infliction of emotional distress. On the other side of the ledger are the protections conferred by the First Amendment and the "fair use" doctrine in copyright law, as well as (for social networking services and other web sites that host third-party content) section 230 of the Communications Decency Act (CDA).

It won't be easy for courts to find the right balance. Rulings that confer overly broad protection to people targeted by deepfakes risk running afoul of the First Amendment and being struck down on appeal. Rulings that are insufficiently protective of deepfake targets could leave people without a mechanism to combat deepfakes that could be extraordinary harmful. And attempts to weaken section 230 of the CDA in the name of addressing the threat posed by deepfakes would create a whole cascade of unintended and damaging consequences to the online ecosystem.

While it remains to be seen how these tensions will play out in the courts, two things are clear today: First, there is already a substantive set of legal remedies that can be used against deepfakes, and second, it's far too early to conclude that they will be insufficient.

Despite this, federal and state legislators, who are under pressure to "do something" about deepfakes, are responding with new legislative proposals. But it is very hard to draft deepfake-specific legislation that isn't problematic with respect to the First Amendment or redundant in light of existing laws.

For example, a (now expired) Senate bill S.3805 introduced in December 2018 would have, among other things, made it unlawful "using any means or facility of interstate or foreign commerce," to "create, with the intent to distribute, a deep fake with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct under Federal, State, local, or Tribal law." Writing at the Volokh Conspiracy regarding S.3805, USC law professor Orin Kerr observed that:

It's already a crime to commit a crime under federal, state, local, or tribal law. It's also already a crime to 'facilitate' a crime—see 18 U.S.C. § 2 at the federal level, and state laws have their equivalents. Plus, it's already a tort to commit a tort under federal, state, local, or tribal law. This new proposed law then makes it a federal crime to either make or distribute a deepfake when the person has the intent to do the thing that is already prohibited. In effect, it mostly adds a federal criminal law hammer to conduct that is already prohibited and that could already lead to either criminal punishment or a civil suit.

State legislators in New York have considered a bill that would prohibit certain uses of a "digital replica" of a person and provide that "for the purposes of the right of publicity, a living or deceased individual's persona is personal property." Unsurprisingly, this raised concerns in the entertainment industry. As a letter from the Walt Disney Company's Vice President of Government Relations stated, "if adopted, this legislation would interfere with the right and ability of companies like ours to tell stories about real people and events. The public has an interest in those stories, and the First Amendment protects those who tell them."

At the end of the day, technological deepfake detection solutions, no matter how good they get, won't prevent all deepfakes from getting distributed. And legal remedies, no matter how effective they might be, are generally applied after the fact. This means they will have limited utility in addressing the potential damage that deepfakes can do, particularly given the short timescales that characterize the creation, distribution, and consumption of digital media.

As a result, improved public awareness needs to be an additional aspect of the strategy for combating deepfakes. When we see videos showing incongruous behavior, it will be important not to immediately assume that the actions depicted are real. When a high-profile suspected deepfake video is published, it will usually be possible to know within days or even hours whether there is reliable evidence that it has been fabricated. That knowledge won't stop deepfakes, but it can certainly help blunt their impact.