Values Embedded in Legal Artificial Intelligence

Harry Surden
University of Colorado at Boulder,
Boulder, CO 80309 USA

Increasingly, governments are using technological systems in the application and administration of law [6], [22]. For example, officials use computer systems to sentence criminal defendants, approve or deny government benefits, predict the location of future crimes, and disallow border entry [6], [22]. In each instance, technology is used to make substantive decisions about law, individual legal rights, or allocation of government resources. Let us refer to any automation system used in the administration or application of law as a "legal technological system." Notably, some of these legal technological systems use machine learning and other artificial intelligence techniques to achieve their goals [13]. Let us refer to those as "legal AI systems."

As legal AI systems become widely deployed, we can draw upon scholars who study the relationship between technology and society. One crucial insight is that technologies can have values subtly embedded in their design [12], [25], [28]. To understand this, consider broadly the development of a new technological system. During that process, engineers (and others) must make design decisions about how that system will operate and what it can or cannot do. These decisions include what features the technology will have, what data to use, the arrangement of a user interface, and the design of the technical and physical architecture, and other operational aspects.

Digital Object Identifier 10.1109/MTS.2022.3147542
Date of current version: 9 March 2022.

Such engineering choices may superficially appear to be value-neutral, given that they are typically made for routine technical, efficiency, or functionality reasons. However, the core idea of "embedded values" is that when a technology becomes widely used in society, technological design choices may end up promoting certain values, or advantaging certain societal subgroups, over others. This can occur even if such values are not intentionally crafted into a technology, but instead have arisen as unforeseen byproducts of decisions made in good faith for functionality reasons.

A brief example of a legal technological design with subtly embedded values will illustrate the point. Around 2000, some U.S. courts began to transition from paper to electronic documents for lawsuits [27]. These court documents included motions and orders (with attachments and exhibits) for legal cases. At the time, the designers of these systems faced a series of engineering decisions, such as whether to make court documents searchable and accessible on the internet, which they ultimately did [27].

The technological choice to put court documents online and make them searchable can be understood as having promoted certain values—such as public accessibility and transparency around litigation—at the expense of other values—such as privacy for litigants. This value shift can be seen by considering how the pre-electronic system provided an implicit layer of privacy for participants. Previously, court documents were in paper form. Therefore, access to any information had to be acquired through a physical visit to the courthouse. Moreover, because paper documents are time-consuming to

0278-0097/22©2022 IEEE

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

search, specific pieces of data within a document trove were difficult to find. Thus, even though sensitive information (e.g., social security numbers or financial information) routinely appeared within court filings that were nominally public, as a practical matter, this information was reasonably protected. It would have been difficult to exploit private data buried amidst hundreds of physical pages located in a distant courthouse.

The move from paper to electronic storage of lawsuit documents represented not only a fundamental change in technological design, but also a change in the values that the system promoted. With the new electronic, networked system, it became easier to remotely access and search for private information previously obscured within paper. Putting the merits of these choices aside, this example illustrates that system design decisions made for engineering, business, efficiency, usability, or other functionality reasons have often had the side effect of subtly (and perhaps unintentionally) advancing certain values (e.g., accessibility of government information) over others (e.g., litigant privacy) when a technology is widely deployed in society.

These embedded-value effects are especially salient when they occur in *legal* technological systems. This is because the issues at stake in legal contexts are often significant, including the possible deprivation of liberty, property, or security. Moreover, because they are often used by governments, legal technological systems with embedded values can impact large portions of the population. Additionally, societal values embedded in legal technological systems can be difficult to detect absent close scrutiny. In particular, such systems can appear to be superficially neutral even when there are value preferences subtly embedded in their architecture. This can be especially pertinent when systems automate decisions about legal rights that were previously made by human decision makers. When human actors make decisions, it is often clear that they are explicitly promoting certain values over others, whereas the social value impacts of widespread, automated decisions may go unnoticed [6], [13].

These embedded value issues can become compounded in legal AI systems—legal technological systems that employ machine learning, formal rule representation, and other artificial intelligence techniques—due to the subtle nature of some of these approaches. Thus, the use of these artificial intelligence systems in legal decision making can raise novel, and perhaps less familiar, issues of embedded values that deserve particular attention compared to AI systems employed in other societal contexts.

Values embedded in technology

Overview of embedded values

This part will introduce some basic themes from the "embedded values" literature before applying these concepts to artificial intelligence within law. Winner [28] was one of the earliest scholars to observe that social values can be subtly embedded within superficially neutral technology. He provided the following (perhaps apocryphal) example of a technological design that privileged certain social groups over others [4], [28]. According to Winner [28], in the 1930s, New York urban planners purposely engineered suburban highways with overpasses that were too low for buses to pass under. He suggested that this design was intended to inhibit low-income and minority citizens, for whom bus was the primary mode of transport, from reaching the suburbs [28]. Winner argued that such a design had the effect of technologically embodying the value preferences of certain societal groups (suburban residents) at the expense of others (low-income and minority urban residents).

Later, Lessig [12] observed the relationship between technological design and social values in the context of the Internet. Lessig [12] recognized that the engineering architecture of a technological system—what the system is designed to allow or disallow—can be thought of as analogous to legal regulation. He noted that while laws explicitly and overtly regulate society, technological design can similarly affect societal behavior or values, but often in less obvious ways. For example, consider the technological design of the Internet and the social impacts of technical choices. At the outset, the designers of the Internet faced numerous engineering decisions—such as whether to provide strong anonymity for users [8], [12]. On the one hand, an engineering constraint built into Internet protocols requiring technological anonymity may have had the effect of fostering certain social values, such as free speech and liberty of conduct in the use of the web. However, such a design would have come at the expense of other values—such as facilitating

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

identification of criminals or curtailing hate speech. Lessig's point was that some engineering choices that seem purely technological in nature can actually be understood as having a socio/political regulatory effect. Because all technological architectures necessarily inhibit some activities and promote others, engineering design can be seen as regulating social behavior implicitly in a manner that is analogous to the explicit government regulation that occurs through law.

More recently, scholars from the algorithmic "fairness, accountability, and transparency (FAT)" literature have revisited these concepts in the context of artificial intelligence and algorithmic governance. For example, Barocas and Selbst [2] have discussed the subtle racial or ethnic biases that can arise in algorithmic decision making. Citron [6] and Kroll et al. [14] (separately) observed the lack of accountability in computer-based decision making that increasingly occurs in government context. Kleinberg et al. [10] identified the fairness and allocation tradeoffs that must inevitably occur in the design of all algorithmic systems.

Overall, Lessig [12], Winner [28], and other scholars raised some fundamental points about technological design and values. First, they illustrated that technological choices made by engineers can have substantive real-world effects upon people's lives, legal rights, and abilities that rival, or even exceed, the impact of overt law and policy choices made by governments. Another, perhaps more important contribution was to challenge a common belief that technological systems, and their design, are somehow neutral with respect to social, political, and other values.

Intentional and unintentional embedding of values

Although technological systems are not value-neutral, it is useful to distinguish between unintentional and intentional embedding of values. This is because the embodiment of values within technology does not necessarily arise due to deliberate manipulation.

To intentionally embed a value is to purposely design a technology to promote one or more values or interests. A good example is "privacy by design." This is the view that information systems should be consciously engineered from the outset with features or processes that promote privacy [22]. For example, a library information system could be purposefully designed to immediately delete data about user borrowing history, rather than storing it persistently [22]. Such a technological design would have the effect of promoting privacy because many privacy violations involve the later acquisition of stored user data records. Although privacy by design is generally thought of as a positive example of purposely structured values, other examples of deliberately embedding values are not so benign. Winner's [28] highway scenario exemplifies how values can be intentionally embedded in technology to harm certain members of society and promote the interests of others.

Perhaps more common, and more pressing however, is the *unintentional* embedding of values. This occurs when certain technological designs are chosen over others for functionality, engineering, efficiency, usability, business, or other legitimate, practical reasons. Unintentional embedding of values occurs when a technology ends up, in practice, advantaging certain values (or social groups) over others as an unforeseen or unwanted byproduct of a chosen design.

For example, consider a system used to approve or deny government benefits (e.g., unemployment benefits). Imagine that to reduce development expenses, the engineers designed the system without an audit trail (i.e., a database that records every action and that can later be used to reconstruct a decision-making process) [6]. While superficially this was a value-neutral engineering decision made for budgetary reasons, such a design can also be understood as unintentionally prioritizing some values over others. The lack of an audit trail may end up promoting the finality of decisions of government officials at the expense of other values, such as appealability for applicants, because the absence of retrievable process record will make it harder for applicants to contest unfavorable benefits decisions [6]. This illustrates how a design decision made for legitimate development purposes can, as a byproduct, unintentionally but significantly promote certain societal values over others when the technology is broadly used.

Such *unintentional* embedding of values is of greater concern for several reasons. First, it is likely much more common than intentional embedding, as embedded values frequently arise only as an unintended side effect of legitimate product design

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

and not the result of deliberate value preferencing. Second, such unintentionally embedded values are often hard to observe. Engineers typically design systems for technical or usability reasons and when that arrangement happens to promote some values over others as a byproduct, that effect can be difficult to detect. Finally, in such cases, there is the legitimate argument that the design choice occurred for reasons completely unrelated to value preferences undermining the claim that any value embedding exists at all.

The point is that there does not have to be a nefarious intent, or any intent at all, for a technology to promote certain values or social groups at the expense of others. Rather any system can come to embody values or biases for completely innocuous reasons. Nonetheless, whether values are intentionally or unintentionally embedded in a technology, the impact on society can be equally impactful when it is broadly deployed.

Aura of technological objectivity

An additional point is that technological systems often have the misleading semblance of objectivity, as compared to human decision makers [1]. This contrast may be particularly evident when such systems are used within the law. When people—such as judges—make decisions, it is obvious that as members of society, they may be preferencing some values over others or be unduly influenced by emotion or bias. By contrast, legal technological systems, particularly those that rely upon data, may superficially appear have an aura of objectivity and neutrality. These systems are nonhuman artifacts, and as such, are not subject to emotion, and they may appear to render objective decisions who outcome is the necessary result of neutral data analysis. For this reason, the value-laden aspects of such technological systems may be overlooked, dismissed, or deferred to, even as they have real-world effects.

For example, consider the systems that judges rely upon to assist in their criminal sentencing decisions. These systems make recommendations based upon analysis of data. However, the output of the system—a formal recommendation report—actually masks an underlying series of subjective judgments on the part of the system designers. These subjective choices include: what data sources to use to build the predictive model, which parts of the chosen data to include or exclude, how to weight that data, what techniques to use to analyze the data, how to validate the system, and what information to emphasize or deemphasize when presenting the analysis to the judge [1]. However, because the recommendation is generated by an automated system using some mechanistic process and appears in stark, computational form, the outcome can have the misleading appearance of near-mathematical objectivity within law.

Because of this aura of mechanistic objectivity, there may be an inclination on the part of judges and others to give more deference to computer-based recommendations, in contrast to comparable human-based assessments. This human tendency to unduly ascribe value neutrality to technological decision making (as compared to similarly situated humans), and to defer to the seeming precision of mathematical and data-based analysis, should be queried closely in the context of technological systems that influence substantive legal outcomes affecting the lives of people.

Building upon the foundational points of embedded values just discussed, the next part will examine areas of concern particularly related to artificial intelligence within law.

Artificial intelligence, law, and embedded values

Recently, systems used within the legal context have begun to incorporate techniques from artificial intelligence including machine learning and formal rule representation. Artificial intelligence raises some novel issues of embedded values than those encountered with previous technologies. This part will survey some common issues of embedded values in legal systems that incorporate artificial intelligence. Designers of legal technological systems should take special care with an eye to the embedded patterns discussed. This is because the law has a unique place in ordering society and in safeguarding the rights and liberties of societal members. When technologies used within the legal system contain subtle value preferences, this can have an outsized impact on society at large.

Biases in data and models

Legal technological systems that incorporate machine learning approaches raise unique values issues. Machine learning refers to a body of techniques from artificial intelligence involving

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

algorithms that are able to detect patterns within data [7]. Often these data-derived patterns are then used to make automated decisions or to engage in predictions. These systems characteristically are able to improve their performance over time—or "learn" on particular tasks—by analyzing additional data. A major requirement for machine-learning is the availability of data—typically large amounts of data—that can be analyzed for useful patterns [7]. The type and quality of the data supplied to a machine learning system is therefore crucial to its functionality.

One of the most important sources of embedded values in legal AI systems come from biases in data. Of major concern: if there are biases in the underlying data used to build a machine learning system, this can lead to biases in how the system performs. These biases may be hard to detect but can produce undesirable, unlawful, or unfair outputs.

For example, as discussed, some U.S. courts use computer models in parole or sentencing decisions for criminal defendants [13]. These systems attempt to predict (among other things) the risk that criminal defendants will commit future crimes. Such predictions are often be based upon general crime data and purport to identify correlations of future criminal behavior. However, often these data sets are heavily based upon recorded police activity. Notably, if the police activity data upon which the machine-learning model was built was itself biased, the predicted outcome might be similarly skewed. For example, imagine that police tended to patrol disproportionally in minority neighborhoods compared to other neighborhoods (thereby skewing the base rate of police interactions with minority populations compared to the populate generally), or if police tended to arrest members of minority groups at a higher rate than nonminority groups, all other things being equal, for similar behaviors. Such biased choices on the part of the police to patrol or arrest at a disproportionate base rate would be reflected in the police recorded data. In other words, the recorded police data would not reflect an objective model of actual crime activity, but would reflect skews introduced by police behavior—patrol, arrest, and data recording practices. This in turn would create a misleadingly high correlation between minority status and risk of offending, which might be embedded into a system analyzing patterns in such police data [13]. Similarly, machine-learning models can detect seemingly neutral correlates—such as zip code—that can become proxies for characteristics that are unlawful to consider, such as race—that might be incorporated by the system.

Importantly, it often is hard to detect when a machine learning system has such biases in its data model because they can become embedded computationally in subtle ways. While such data-induced biases are of concern in computing generally, they are particularly problematic in legal systems given the substantial real-world effects that they can have on the legal rights of individuals. To the extent that such systems are used in the application of law, designers should be particularly attentive to the possibility that the data upon which their model was built might be systematically skewed in some way that might preference or inhibit some social subgroups.

Inscrutable and uninterpretable models

Some artificial intelligence models are difficult to interpret and understand. This idea goes by various names, including the "intelligibility," "comprehensibility," or "inscrutability" problems [13], [17]. The general idea is the following: When we create a computer system, we often need to know why the system made a particular decision that it did. However, while all systems encode their decision-making processes in some sort of computer model, some models are more understandable by people than others. For example, certain rules-based artificial intelligence systems can provide a logical, step-by-step analysis, in human-readable form, as to why they took the particular decision that they did. Similarly, human written source code tends to be comparatively easy to understand, as programmers can systematically proceed step by step through the instructions to understand why a piece of software acts the way that it does.

By contrast, some machine-learning techniques produce extremely complex computer models whose underlying logic can be very difficult, if not impossible, for humans to inspect and comprehend. For example, neural networks—particularly deep learning neutral network systems—have proved very adept at automating certain tasks. Notably, however, neural networks tend to encode their patterns in models that are notoriously difficult to understand. In many cases, neural networks are able to produce highly accurate results on complex tasks using an

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

underlying mechanism that is not interpretable by people. Often even the programmers who created a neural network do not understand how it works, nor why it reached the decision that it did. In sum, it is possible to have a very effective machine-learning model—in the sense that it makes very accurate, useful decisions on complex tasks—but whose inner workings are comparatively difficult, if not impossible, for people to understand and interrogate.

Such inscrutable technological models may be particularly problematic within law. Basic legal principles require that legal decision makers be able to explain why they came to the decisions that they did. Articulated rationale is a central tenet of legal decision making, particularly when decisions involve the deprivation of liberty or property. However, to the extent that legal officials are assisted by artificial intelligence systems that have core interpretability limitations, such articulable rationales may not be possible, undermining central legal norms. Moreover, as previously described, such uninterpretable mathematical models may further mask underlying, and undesirable, biases that may be difficult to detect by human inspection.

However, while it is generally desirable to make machine-based legal decisions more interpretable, it is important to acknowledge that such technical changes will not solve underlying structural–legal problems where they exist. In some cases, it may make them worse if not done carefully and in good faith. As discussed earlier, machine decisions often reflect a superficial veneer of value neutrality, when in fact the technology can contain subtle, embedded value preferences. The concern is that improving a technical aspect of a legal AI system may leave the false impression that a deeper, underlying legal–societal issue has been resolved and no longer requires attention.

Formalizing law and rules-based systems

A different class of artificial intelligence techniques, grounded in computer logic and rules, provides distinct embedded values issues within law. This rules-based approach typically involves the formal representation of laws on a computer system using computer logic [13]. In such an approach, computer programmers, lawyers, and others examine particular laws and attempt to translate them into a set of comparable rules that a computer can follow, while attempting to preserve the underlying logic and meaning of the laws. The translation of law and legal relationships into computer structures, such as rules and ontologies, is sometimes called formalization.

Many such systems aim to aid in legal decision making or to allow the assessment of legal outcomes. For example, in the United States, software systems, such as Turbotax, assist citizens in filing their personal income taxes. These software systems attempt to formally model a portion of U.S. personal income tax laws; they aim to replicate the underlying logic of the laws in computer processable form and allow for computation of tax liability under the law [26].

A major issue is that the very act of "translating" laws into computer rules actually masks a series of subjective and contestable decisions about the meaning and scope of the law. Many applications of law involve unpredictability—uncertainty about which laws do, or do not, apply to a given situation, how these laws are to be interpreted, and once interpreted what the outcome will be when applied to particular facts. In a typical legal scenario, there are multiple plausible interpretations of a given law, each with slightly different meanings and scope, and all which are reasonable. One of the major roles of legal officials—such as judges—is to resolve these uncertainties in applying laws in particular circumstances, taking into account considerations such as the text, meaning, and purpose of the law, and competing public policies. Society often does not often know a definitive answer about such legal uncertainty until a legal official makes a binding, final determination, electing one set of possible arguments and interpretations over others.

By contrast, the process of "translating" a law for a computer system necessarily involves a series of judgment calls about the meaning, content, and applicability of the law (and other legally uncertain issues) on the part of the translator. Thus, to formalize a law in as a series of computer rules is to commit to one particular legal interpretation over others. This process involves implicitly choosing one set of contestable arguments about the meaning and scope of the law over other plausible readings. Problematically, it may not be obvious, even to the programmers who engage in such formalization, that translating a law into comparable computer code actually involves an implicit set of subjective, and perhaps value-laden, interpretive choices.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

A primary issue is that in rules-based legal systems, such value judgments about the law become fixed in computer code. Moreover, these choices become embedded in technology where their impact can be magnified when the software is distributed widely. For example, the software model created by Turbotax reflects a series of subjective interpretations about the meaning of the U.S. income tax code made by the employees of the Intuit corporation. Their interpretive choices about the meaning of the personal income tax laws become embedded in the software itself. The impact of these interpretations become magnified when Turbotax's software is distributed to millions of users worldwide who implicitly adopt this value-laden model.

In sum, the formal representation of law in a legal technological system may reflect a series of subtly encoded values and subjective decisions that may be difficult to detect. Moreover, when these values are embedded as software rules, the impact of these subjective judgments can become magnified when the software model is distributed and adopted broadly.

Removing prior structural constraints

A final way in which values can be subtly embedded in legal technological systems has to do with the reduction of existing technological constraints by artificial intelligence.

Observe that sometimes a new technology will make an activity that used to be difficult, suddenly much easier or less expensive, or significantly reduce transaction costs. In some instances, simply making some activity technologically easier to do than it was before may be subtly value laden. Consider the court record example from the earlier discussion. In the era of paper documents, court filings were more difficult to access. The constraints of paper technology meant that finding private information in a large number of public court documents was prohibitive. Accessing court documents in that pre-electronic era typically required physical entry to the courthouse and then cumbersome, manual searching of volumes of papers to find particular private information. By digitizing court documents and putting them online, the activities of accessing and searching for private data within voluminous court filings, as an unintended byproduct, became dramatically less difficult.

In the prior technological era, we can think of paper technology as having *implicitly* protected the privacy of litigants by providing a structural constraint. The subsequent digital technology removed this implicit constraint, and as a side effect, diminishing litigant privacy. A similar effect has occurred with the digitization of other government records such as deeds, voter registration documents, and campaign contributions, which were always nominally public, but in the prior paper technological era, practically difficult to access and analyze. Digitizing these public documents, and putting them online has as a byproduct, reduced privacy, because this technology makes it much easier to access previously disparate public data about citizens and cross-reference, aggregate, and centralize this information (e.g., linking an individual's campaign contribution, housing purchase, and vehicle registration public records) than in the past.

Overall, any new technology that makes some activity substantially easier than it used to be may have the effect of unintentionally diminishing some value (such as privacy) which was implicitly protected by the constraints inherent in the prior technological era. The central idea is that sometimes technical cost, inefficiency, or "friction" can be understood as playing a non-obvious implicit functional role in protecting societal values. When a new technology, such as machine learning removes or reduces frictions or inefficiencies that also were serving a (perhaps unrecognized) functional role in fostering values in the previous technological context, the diffusion of the technology be understood as widely shifting certain embedded societal values (e.g., efficiency and inferring correlations) at the expense of others (e.g., privacy and anonymity).

This can be seen within the legal domain. Artificial intelligence techniques such as machine learning make particular activities that used to be difficult—such as detecting patterns in large data sets or aggregating previously disparate data, much easier. For instance, it is now often possible to probabilistically infer private information about a person that has not been publicly revealed—such as sexual orientation—simply by scanning publicly available data, such as social network connections, and engaging in statistical analysis [28]. Such capabilities may be particularly problematic in law, where machine-learning techniques now make it significantly easier to surreptitiously detect correlates (e.g., zip codes) for categories that may be unlawful to consider directly (e.g., race and sexual orientation).

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

Efforts to address embedded values issues

It is important to observe that there are promising developments that may help mitigate some of the problems just described. There is current research aimed at addressing multiple issues related to data bias and embedded values. For example, one line of research aims to make the internal structure of inscrutable computer models more understandable to people [21]. Several research groups have developed computers systems that can analyze "black-box" computer algorithms and provide basic explanations for automated decisions that are understandable to people [15], [18], [21]. Other research programs are focusing upon detecting bias in data sets, or attempting to ensure that machine learning models produce fairer and more proportionate results [1], [29]. Moreover, note that it is not just computers systems that are subject to embedded biases and values. Human decision makers, such as judges, have long been subject to biases in their judgments that have historically difficult to detect. Machine-learning research offers promise in detecting biases in judges and other human decision makers as well [5].

Observe, however, that this line of inquiry is relatively new and is currently in its formative stages outside of law [24]. As such, relatively few of these bias-mitigating developments appear to have made their way into the technological systems that are just now entering widespread use in the legal context. However, the incorporation of some of these research findings into legal technological systems could, in the future, help to mitigate, or render more transparent, some of these embedded value problems. The creators of legal technological systems should pay close attention to developments in emerging area of bias-mitigating and values-aware technical research.

Technological systems that use artificial intelligence are increasingly being used in the application of law. Such systems can contain values subtly embedded in their technological design. This observation becomes particularly important in the context of law, given the significant issues at stake, including loss of liberty, property, or rights. Legal technological systems that employ artificial intelligence require special care and awareness in development, as the use of artificial intelligence can raise specific issues of embedded values that may be impactful but hard to observe. New technological developments that can address some of these issues, as well as vigilance on the part of developers, may help mitigate certain problems. However, as previously discussed, it is important not to frame these topics solely as technical issues to be solved. In many cases, the legal AI issues merely reflect deeper underlying social and structural issues that technical solutions cannot fully address. ■

References

[1] A. Agarwal et al., "A reductions approach to fair classification," in *Proc. FAT ML*, 2017, pp. 60–69.

[2] S. Barocas and A. Selbst, "Big data's disparate impact," *California Law Rev.*, vol. 671, p. 104, Sep. 2016.

[3] J. O. Berger and D. A. Berry, "Statistical analysis and the illusion of objectivity," *Amer. Sci.*, vol. 76, pp. 159–165, Apr. 1988.

[4] B. Joerges, "Do politics have artefacts?" *Social Stud. Sci.*, vol. 29, pp. 411–431, Jun. 1999.

[5] R. Vunikili et al., "Analysis of vocal implicit bias in SCOTUS decisions through predictive modelling," *Proc. Exp. Linguistics*, pp. 121–124, Dec. 2018.

[6] D. Citron, "Technological due process," *Washington Univ. Law Rev.*, vol. 85, no. 6, p. 1249, 2009.

[7] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, pp. 78–87, Oct. 2012.

[8] I. Goldberg and D. Wagner, "TAZ servers and the rewebber network: Enabling anonymous publishing on the world wide web," *First Monday*, vol. 3, no. 4, 1998. [Online]. Available: https://doi.org/10.5210/fm.v3i4.586

[9] W. Hartzog and F. Stutzman, "Obscurity by design," *Washington Law Rev.*, vol. 88, p. 385, Jun. 2013.

[10] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innov. Theor. Comput. Sci. Conf.*, 2017, pp. 43:1–43:23.

[11] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, 2013.

[12] L. Lessig, *Code 2.0*. New York, NY, USA: Basic Books, 2006.

[13] N. Love and M. Genesereth, "Computational law," in *Proc. 10th Int. Conf. Artif. Intell. Law*, 2005, pp. 205–209.

[14] J. Kroll et al., "Accountable algorithms," *Univ. Pensylvania Law Rev.*, vol. 165, p. 633, Nov. 2016.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.

[15] J. McAuley and J. Leskovec, “Hidden factors and hidden topics: Understanding rating dimensions with review text,” in *Proc. RecSys*, 2013, pp. 165–172.

[16] R. Michalski and Y. Kodratoff, “Research in machine learning: Recent progress, classification of methods, and future directions,” in *Machine Learning: An Artificial Intelligence Approach*, vol. 3. Elsevier–Morgan Kaufmann Imprint 2014, p. 6.

[17] C. Molnar, *Interpretable Machine Learning*. Lulu 2020.

[18] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2017.

[19] C. O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[20] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA, USA: Harvard Univ. Press, 2015.

[21] M. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in *Proc. KDD*, 2016, pp. 1135–1144.

[22] A. Roth, “Trial by machine,” *Georgetown Law J.*, vol. 1245, p. 104, Mar. 2016.

[23] P. Scharr, “Privacy by design,” *Identity Inf. Soc.*, vol. 3, no. 2, pp. 267–274, 2010.

[24] A. Selbst et al., “Fairness and abstraction in sociotechnical systems,” in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 59–68.

[25] D. H. Guston and J. Stilgoe, “Responsible Research and Innovation,” in *Handbook of Science and Technology Studies*, U. Felt et al., Eds., 4th ed. Cambridge, MA, USA: MIT Press, 2017, pp. 853–880.

[26] H. Surden, “The variable determinacy thesis,” *Columbia Law Technol. Rev.*, vol. 12, no. 1, p. 100, 2011.

[27] H. Surden, “Structural rights in privacy,” *SMU Law Rev.*, vol. 60, pp. 1605–1629, 2008.

[28] L. Winner, “Do artifacts have politics?” *Daedalus*, vol. 109, no. 1, pp. 121–136, 1980.

[29] Z. Zhang and D. B. Neill, “Identifying significant predictive bias in classifiers,” 2016, *arXiv:1611.08292.*

Harry Surden is a Professor of law at the University of Colorado at Boulder, Boulder, CO, USA, and an Affiliated Faculty Member at Stanford University’s CodeX Center for Legal Informatics, Stanford, CA, USA. His scholarship focuses upon legal informatics, artificial intelligence and law (including machine learning and law), legal automation, and issues concerning self-driving/autonomous vehicles. Surden has a bachelor’s from Cornell University, Ithaca, NY, USA, and a JD from Stanford University.

■ Direct questions and comments about this article to Harry Surden, University of Colorado at Boulder, Boulder, CO 80309 USA; hsurden@colorado.edu.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 28,2024 at 18:04:44 UTC from IEEE Xplore. Restrictions apply.