Values Embedded in Legal Artificial Intelligence

U of Colorado Law Legal Studies Research Paper No. 17-17

Harry Surden

Values Embedded in Legal Artificial Intelligence

Harry Surden
Associate Professor of Law
University of Colorado Law School
Boulder, CO 80309
hsurden@colorado.edu

ABSTRACT

Technological systems can have values embedded in their design. This means that certain technologies, when they used, can have the effect of promoting or inhibiting particular societal values over others. Although sometimes the embedding of values is intentional, often it is unintentional and, in either case, when it occurs it can be difficult to observe or detect. The embedding of values in technological systems becomes increasingly significant when these systems are used in the application of law. This article explores embedded values challenges posed by legal technological systems that use artificial intelligence.

Keywords

Technological values, artificial intelligence, law, machine learning, embedded values.

1. INTRODUCTION

Increasingly, technological systems are being used in the application and administration of law. [3] [13] For example, government officials use computer systems to sentence criminal defendants, approve or deny government benefits, predict the location of future crimes, and disallow border entry. [3] [13] In each instance, technology is used to make substantive decisions about law, individual legal rights, or allocation of government resources. Let us refer to a system used in the administration or application of law as a "legal technological system." Notably, some of these systems use machine learning, and other artificial intelligence techniques, to achieve their goals. [9]

As scholars have observed, technological systems can have values embedded in their design. [8] [18] [15]. In developing a new technology, engineers must make a series of design decisions about how the technology will operate and what it can or cannot do. These decisions include what features and limitations the technology should have, what data to use and how to analyze and present it, the arrangement of a user interface, and the design of the technical and physical architecture. Such engineering choices may superficially appear to be value-neutral. However, the core idea of "embedded values" is that when a technology becomes widely used, technological design choices may end up promoting certain values, or advantaging certain societal subgroups, over others. Importantly, such value-preferencing is often not intentional, but rather a byproduct of design decisions made for technical, efficiency, or functionality reasons.

A brief example of a 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. [17] These court documents included motions and orders (with attachments and exhibits) for particular legal cases. The designers of these document systems faced a series of engineering decisions, such as whether to make court documents searchable or accessible on the Internet. [17] The technological choice to make court documents online and searchable could be understood as promoting certain values – such as public accessibility and government transparency. However, such a design choice could also be understood as coming at the expense of other values – such as privacy for litigants. The information in these court documents, which had previously been accessible through physical access in paper form, and which had not been remotely searchable, underwent a fundamental change in technological design. The new features made it easier for third parties to remotely access and search for private information buried within court documents (e.g. social security numbers of litigants located within court filings). Putting the merits of this choice aside, the point is that system design decisions made for engineering, business, efficiency, usability, or other functionality-based reasons can often have the side-effect of subtly advancing certain social values (e.g. accessibility of government information) over others (e.g. litigant privacy) when a technology is deployed.

It is especially important to observe when values are embedded in legal technological systems. One reason is that the issues at stake in legal contexts are often significant, including the possible deprivation of liberty, property, or security. Additionally, values embedded in technological systems can be difficult to detect absent close scrutiny. In particular, systems can appear to be superficially neutral, when it fact there may be values subtly embedded in the technological architecture. Such observations are especially pertinent where systems partially (or fully) automate decisions about legal rights that were previously made by human decision-makers who were understood as engaging in explicit value weightings. [3] [9] Finally, legal technological systems that use machine-learning, formal rule representation, and other artificial intelligence techniques raise novel, and perhaps less familiar, issues of embedded values that deserve particular attention.

2. Values Embedded in Technology

2.1.1 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. Langdon Winner was one of the earliest scholars to observe that values can be embedded in technological design. Winner provided the following (contested) example of design privileging certain social groups over others. [18] [2] According to Winner, in the 1930's, New York urban planners purposely engineered suburban highways with overpasses that were too low for busses to pass under. [18] This design was intended to inhibit low-income and minority citizens, for whom bus was the primary mode of transport, from reaching the suburbs. [18] 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).

More recently, Lawrence Lessig observed the relationship between technological design and social values in the context of the Internet. Lessig noted 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. [8] For example, the designers of the Internet faced a series of engineering decisions – such as whether to provide strong anonymity for users. [5] [8] An engineering choice to require strong 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 decision might have come at the expense of other social values - such as facilitating identification of criminals or curtailing hate speech. Lessig's point was that engineering choices about the architecture of systems such as the Internet could be understood as quasi-regulatory. Because certain technological architectures inhibit some activities and promote others, engineering design can be understood as orchestrating behavior in a manner analogous to explicit government lawmaking.

Overall, Winner, Lessig, and other scholars of this genre raised some fundamental points about technological design, social values, and preferencing. First, they illustrated that technological choices made by engineers can have substantive real-world effects upon people's lives, rights, and abilities. In some cases, the impact of these design choices are commensurate to the overt law and policy choices made by lawmakers and regulators. Another, perhaps more important contribution was to challenge common a belief that technological systems are value-neutral.

2.1.2 Intentional and Unintentional Embedding of Values

It is useful to distinguish between unintentional and intentional embedding of values because the embodiment of values in technology is not necessarily purposeful.

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. [13] For example, a library information system could be purposefully designed to immediately delete data about user borrowing history, rather than storing it persistently. [13] Such a technological design would have the effect of promoting privacy, because many privacy violations involve the acquisition or release of stored user data records. However, the purposeful embedding of values is not always as benign as in privacy by design. Winner's highway scenario is an example of intentionally embedding of values to promote the interests of one societal group to the detriment of others. [18]

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

For example, consider a government system for making decisions about the approval of denial of government benefits. Imagine that in order to reduce development expenses, the engineers of the system designed it without an audit trail (a database that records every action taken, often with reasons why, that can later be used to reconstruct a decision-making process). [3] While superficially 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 values such as finality of decisions of government officials, at the expense of others, such as appealability for applicants. [3] Because the system has no audit trail, there will be no retrievable decision record that appellants can use, and it will be harder to contest unfavorable benefits decisions.

The 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, and not the result of deliberate value preferencing. Second, such unintentionally embedded values are often hard to observe. When engineers design systems for technical or usability reasons and that arrangement only happens to promote some values over others as a byproduct, the 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 technology can embody value preferences or biases for completely innocuous reasons unrelated to whether anyone intended for those value preferences to be there. However, values unintentionally embedded in technology can be just as impactful as those intentionally embedded, when the technology becomes widely deployed, particularly when deployed in legal technological systems.

2.1.3 The Aura of Technological Objectivity

An additional point is that legal technological systems can have the semblance of value-objectivity even when this is not the case. [1] One reason is the contrast between human and mechanical decision-making. When people – such as judges - make decisions, it is obvious that, as humans, 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 appear have a superficial aura of objectivity and neutrality because they are non-human artifacts. 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 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 about what data to use, include or exclude, how to weight the data, and what information to emphasize or deemphasize to the judge. [1] Because the recommendation is generated by an automated system, using some mechanistic process, the outcome can have the appearance of creating an illusion of computed mathematical objectivity within law.

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, given the aura of mechanistic objectivity surrounding computer-generated, analytics-based, analyses. This human tendency to unduly ascribe value neutrality to technology, and to defer to mathematical analysis, should be queried closely in the context of technological systems that influence substantive legal outcomes.

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

3. Artificial Intelligence, Law, and Embedded Values

Recently, legal technological systems have incorporated techniques from artificial intelligence, including machine-learning and formal rule representation. Artificial intelligence raises some novel, significant, but less traditional patterns of embedded values. This part will survey some of common issues of embedded values in legal systems that incorporate artificial intelligence. Designers of such systems should take special care with an eye to these patterns.

3.1.1 Biases in Data and Models

Legal technological systems that incorporate machine learning raise unique embedded values issues. Machine learning refers to a body of techniques from artificial intelligence involving algorithms that are able to detect patterns within data. [4] 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 assessing data. A major requirement for machine-learning is the availability of data – typically large amounts of data – that can be analyzed for useful patterns. [4] The type and quality of the data supplied to a machine learning system is thus crucial to its functionality.

One of the most important sources of embedded values in such legal systems are biases in data. Of major concern: if there are biases in the underlying data-set used to build a machine learning system, this can lead to hard-to-detect biases in the system itself that 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.[9] Some of these systems attempt to predict risk of reoffending based upon past crime data by identifying correlations between arrests and criminal behavior. Often these data-sets are based upon recorded police activity. However, if the police activity upon which the machine-learning model was built, was itself biased in some way, the predicted outcome might be similarly skewed. For example, imagine that police tended to arrest minority groups at a higher rate than non-minority groups all other things being equal. Such biased activity could create a disproportionately high correlation between minority status and risk that might be unintentionally incorporated within the system. [9] 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 embedded 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.

3.1.2 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" problem.[9] 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 decisions 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 proven 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 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.

3.1.3 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. [9] 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.[16]

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.

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 world-wide 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.

3.1.4 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 court house, 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.

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 publically revealed – such as sexual orientation – simply by scanning publically available data, such as social network connections, and engaging in statistical analysis. [18] 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, sexual orientation).

4. Conclusion

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.

5. REFERENCES

[1] James O. Berger & Donald A. Berry, *Statistical Analysis and the Illusion of Objectivity*, 76 AMERICAN SCIENTIST 159–165 (1988).

[2] Bernward Joerges, *Do Politics Have Artefacts?*, 29 Social Studies of Science 411–431 (1999) (Contesting Langdon Winner's bridge example)

[3] Citron, D, *Technological Due Process,* Wash. U. L. Rev. (2009).

[4] Pedro Domingos, A Few Useful Things to Know about Machine Learning, 55 COMMUN. ACM. 78–87 (2012).

[5] Goldberg, I., And Wagner, D. *Taa servers and the Rewebber network: Enabling Anonymous Publishing on the World Wide Web*. First Monday 3, 4 (1998).

[6] Hartzog, W and Stutzman, F, *Obscurity by Design, Washington Law Review* (2013)

[7] Kosinskia, M, et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*.

[8] Lessig, L. *Code 2.0,* Basic Books (2006).

[9] Love, N. & Genesereth, M., *Computational Law*, Proceedings of the 10th Internationl Conference on Artificial Intelligence and Law.

[10] Michalski, R & Kodratoff, Y *Research in Machine Learning: Recent Progress, Classification of Methods, and Future Decisions*, *in* Machine Learning: An Artificial Intelligence Approach, Volume III 3, 6 (2014).

[11] O'Neil, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown (2016)

[12] Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press (2015).

[13] Roth, A, *Trial By Machine*, 104 Georgetown Law Journal 1245(2016).

[14] Peter Scharr, *Privacy by Design*, Identity in the Information Society (2010).

[15] Stilgoe, J and Guston, D , *Responsible Research and Innovation* (2016)

[16] Surden, H. *The Variable Determinacy Thesis*, Columbia Law and Technology Review (2011).

[17] Surden, H., *Structural Rights in Privacy, SMU Law Review,* (2008).

[18] Winner, L., *Do Artifacts Have Politics?,* Daedalus, 109 (1980)