Masthead Logo

Santa Clara High Technology Law Journal

Volume 35 | Issue 4

Article 2

4-20-2019

ALGORITHMS AND HUMAN FREEDOM

Robert H. Sloan

Richard Warner

ALGORITHMS AND HUMAN FREEDOM

By Robert H. Sloan and Richard Warner [†]

Predictive analytics such as data mining, machine learning, and artificial intelligence drive algorithmic decision making. Its "all-encompassing scope already reaches the very heart of a functioning society". Unfortunately, the legal system and its various tools developed around human decisionmakers cannot adequately administer accountability mechanisms for computer decision making. Antiquated approaches require modernization to bridge the gap between governing human decision making and new technologies.

We divide the bridge-building task into three questions. First, what features of the use of predictive analytics significantly contribute to incorrect, unjustified, or unfair outcomes? Second, how should one regulate those features to make outcomes more acceptable? Third, how can one ensure that the use of predictive analytics sufficiently respects human freedom? We divide the bridge-building task into three questions. First, what features of the use of predictive analytics significantly contribute to "incorrect, unjustified, or unfair" outcomes? Second, how should one regulate those features to make outcomes more acceptable? Third, how can one ensure that the use of predictive analytics sufficiently respects human freedom? You are not free when you are subject to the arbitrary will another, and predictive analytics is no exception. It violates your freedom when it pushes you down an arbitrary and capricious path.

We answer the first question by "profiling" uses of predictive analytics. We adapt the idea of profiling people. A profile of a person is a summary of characteristics relevant to evaluating and predicting the person's behavior. Our profile consists of five features that significantly affect the extent to which a system will yield "incorrect, unjustified, or unfair" decisions. We answer the second question by explaining how to control predictive systems by regulating the features the profile identifies. Along with others, we propose that a government agency regulate the use of predictive systems. The novel feature of our approach is the use of legal regulation to unify consumer demand in ways that create a type of norm extensive studied in game theory, a coordination norm.

[†] Robert H. Sloan is the Professor and Head, Department of Computer Science, University of Illinois at Chicago. Partially supported by National Science Foundation Grant No. DGE-1069311. Richard Warner is a Professor of Law, Chicago-Kent College of Law.

CONTENTS

INTRODUCTION

Predictive analytics is the use of use of mathematical, statistical, and artificial intelligence techniques for classification and prediction.[1] We will use *predictive systems* for systems using the techniques of predictive analytics. Such systems have already yielded significant benefits and we take it for granted that they will continue to do so and are in part for that reason well-entrenched.[2] Indeed, their "all-encompassing scope already reaches the very heart of a functioning society."[3] We focus on a subset of predictive systems — those used to predict *individual human actions*. In this case especially,

> [T]he accountability mechanisms and legal standards that govern decision processes have not kept pace with technology. The tools currently available to policymakers, legislators, and courts were developed primarily to oversee human decisionmakers. Many observers have argued that our current frameworks are not well adapted for situations in which a potentially incorrect, unjustified, or unfair outcome emerges from a computer. Citizens, and society as a whole, have an interest in making these processes more accountable.

---

[1] This characterization of predictive analytics is sufficient for our purposes. There is no agreement on the precise meaning of predictive analytics and related terms like data mining, machine learning, artificial intelligence, neural nets, and deep learning. *See, e.g.*, STEVEN FINLAY, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR BUSINESS: A NO-NONSENSE GUIDE TO DATA DRIVEN TECHNOLOGIES 5-16, 27-59 (2nd ed. 2017) (distinguishing and discussing relationships among machine learning, predictive analytics, data mining, artificial intelligence, neural nets, and deep learning); VIJAY KOTU & BALA DESHPANDE, PREDICTIVE ANALYTICS AND DATA MINING: CONCEPTS AND PRACTICE WITH RAPIDMINER 13-15 (2014) (discussing relations between data mining and predictive analytics).

[2] *See* FOSTER PROVOST & TOM FAWCETT, DATA SCIENCE FOR BUSINESS: WHAT YOU NEED TO KNOW ABOUT DATA MINING AND DATA-ANALYTIC THINKING 1 (2013).

[3] ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE 293 (2016). Siegel identifies 182 different types of use. *Id.* at 191. A short list of examples includes the extension of credit, marketing and advertising, judicial sentencing and parole decisions, searching travelers, auditing taxpayers, police scrutiny of individuals and neighborhoods, welfare and financial aid, public health decisions, employee hiring, visa decisions, political campaign decisions, business planning and supply chain management, call center treatment, employee scheduling, evaluation of teachers, and ranking of the value of customers for differential treatment. *See generally* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 662-63 n. 97 (2017); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); FINLAY, *supra* note 1, at 9 ("Today, machine learning is being applied to a huge range of problems. In fact, almost any aspect of life that involves decision making in one form or another.").

> If these new inventions are to be made governable, this gap must be bridged.[4]

We divide the bridge-building task into three parts. What features of predictive systems significantly contribute to unfair or otherwise objectionable outcomes? How should one regulate those features? Lastly, how can one ensure that predictive systems sufficiently respect human freedom? The third question requires some explanation. The essential point is that you are not free when you are subject to the arbitrary will of another. Predictive systems are no exception. Predictive systems violate your freedom when they push you down an arbitrary and capricious path, and — importantly for our purposes — they also violate your freedom when you are left with no practical alternative but to submit to the decisions without knowing whether there are adequate reasons for them, reasons that at a minimum show that the decisions are not arbitrary and capricious. You are not free if you are subject to the will of another and denied knowledge of whether that will is arbitrary and capricious. Respecting human freedom requires meeting, or at least sufficiently closely approximating, the following *Knowledge Condition*: those subject to decisions of another are able with reasonable effort to know that there is an adequate justification for the decisions.[5]

We answer the first question about objectionable features by "profiling" predictive systems. We adapt the idea of profiling people. A profile of a person is a summary of characteristics relevant to evaluating and predicting the person's behavior.[6] Our predictive systems profile consists of five computational design features that significantly affect the extent to which a system will yield unfair or otherwise objectionable decisions. A *computational design feature* is a feature that is part of the computational strategy for generating an

---

[4] Kroll et al., *supra* note 3, at 636 (footnotes omitted).

[5] *Compare* Devan R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide To Algorithms and the Law*, 31 HARV. J. L. TECH. 1, 9 (2017) (Desai and Kroll formulate the requirement in terms of dignity: "dignity [i.e. our freedom] requires those who are subject to such a process know or understand what reasons are behind a decision." The difference between their formulation and ours is largely verbal.), *with* JERRY L. MASHAW, REASONED ADMINISTRATION AND DEMOCRATIC LEGITIMACY: HOW ADMINISTRATIVE LAW SUPPORTS DEMOCRATIC GOVERNMENT 177 (2018). (noting that "reason-giving is critical to treating individuals as free moral agents subject to legitimate coercion only to the extent that appropriate reasons can be given for restricting their freedom of action"). We assume that respecting human freedom is intrinsically desirable, but also that one can in some cases justify restricting it on consequentialist grounds.

[6] *See, e.g.*, JOE NAVARRO & MARVIN KARLINS, WHAT EVERY BODY IS SAYING: AN EX-FBI AGENT'S GUIDE TO SPEED-READING PEOPLE (2008) (detailing profiling based on nonverbal cues).

intended result. The features in the profile are non-technical ones accessible to those unfamiliar with predictive analytics.[7]

We answer the second question by explaining how to regulate the features the profile identifies. Along with others, we suggest the Federal Trade Commission as the regulatory agency.[8] Our novelty is the use of FTC regulation to create a type of norm extensively studied in game theory, a coordination norm. We suggest using such norms to coordinate consumer or seller activity in ways that create market incentives to minimize unfair or otherwise objectionable decisions. Our answer to the second question is the basis for our answer to the third. We appeal to coordination norms to explain how to meet the *Knowledge Condition*. Our second and third answers are sketches — sufficient for our purposes, but not full explanations.[9] We offer them as a way to begin a discussion about how to use an explicit profile to regulate predictive systems in ways that avoid objectionable consequences while also meeting the *Knowledge Condition*. We confine our attention to commercial contexts. Government surveillance raises related but distinct concerns. Much of what we say remains relevant, however.

Section I identifies features of predictive systems that lead to results that are both unfair and offenses to human freedom. Section II raises the question of the extent to which current predictive systems exhibit those features. It answers the question by constructing a profile of predictive systems. Section III shows how coordination norms can facilitate the fulfillment of the *Knowledge Condition*. Section IV explains how to create norms governing predictive system that ensure that predictive systems do not exhibit the objectionable features

---

[7] In this way, our profile differs from prior scholars' characterizations of policy issues of the sort Ryan Calo usefully offers. *See* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 407-9 (2017). The work closest to our proposal is David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669-70 (2017). We share with Lehr and Ohm a computational design focused characterization of predictive systems based on a thorough technical understanding of predictive systems. We also share similar views on data collection and preparation, problem definition, and the use of proxies. We differ in excluding from our profile more technical details of model selection, training, and evaluation. This allows us to give a profile that is fully general (as Lehr and Ohm note, some of their characterizations apply only to certain types of systems) and accessible to those not well versed in technical details. The two characterizations are complementary.

[8] *See, e.g.*, Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 20 (2014) (advocating for the FTC to oversee the use of data mining and predictive analytics in credit-scoring systems).

[9] For a detailed explanation, *see* ROBERT H. SLOAN & RICHARD WARNER, THE PRIVACY FIX: HOW TO PRESERVE PRIVACY IN THE ONSLAUGHT OF SURVEILLANCE (forthcoming 2020).

Section I identifies while also ensuring that the *Knowledge Condition* is fulfilled.

I. WHY IS PREDICTIVE ANALYTICS SO TROUBLING?

Predictive systems can lead to objectionable outcomes, but so can governments and markets. What makes predictive analytics particularly worrisome? We answer by identifying four features of predictive analytics that are cause for serious concern.

A. Alien Intelligence

We identify those features with a thought experiment. Imagine aliens from outer space land. They are beneficent (or at least well intentioned) aliens who come in peace. One of their first acts is to provide their human hosts with a collection of predictive analytics and artificial intelligence systems that predict future individual human actions. Call them collectively AI, for alien intelligence. Humans — businesses, governments, and individuals — embrace the program, and many (humans) propose using AI systematically in the widest possible range of contexts as a basis for decisions based on its classification and prediction. Would that be a good idea? That depends on the features of AI.

We assume the AI has the following features (for convenience, we describe AI as making decisions even though it is the humans using it who decide):

- *Low accuracy*: AI's classifications and predictions are more accurate than human predictions (even with human predictive systems). But, like human systems, it still has a high error rate in a significant number of cases.[10]
- *Opacity*: Humans have no explanation or understanding of why AI predicts what it does. Even the best human computer science experts find large parts of AI

[10] *See, e.g.*, Andrew Guthrie Ferguson, *The Police Are Using Computer Algorithms to Tell if You're a Threat*, TIME (Oct. 3, 2017), http://time.com/4966125/police-departments-algorithms-chicago/ (addressing how the racial bias, history of police interaction, and other factors can skew an algorithm that calculates and individual's numerical "threat score" in Chicago.). *See also* Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347, 363 (2016) (Chicago Strategic Subject List ("SSL") algorithm, meant to identify potential shooting perpetrators and victims "identified less than 1 % of homicide victims"); *Strategic Subject List*, CITY OF CHICAGO (Dec. 7, 2017), https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np (the seeding data set for the aforementioned Chicago SSL algorithm).

completely opaque. It appears to involve unknown programming and statistical techniques.

- *Broad-based predictions*: What happens in virtually any area of one's life may serve as input to classifications and predictions affecting virtually any other area (what you pay for insurance, whether you get hired, what schools you get in to, and so on).
- *Data feedback without error correction*: AI's decisions affect what happens to people in the future, and that information feeds back into AI as input for subsequent decisions. AI does not, however, have mechanisms that detect and correct its errors. Data feedback without error correction combines with broad-based predictions to create mistaken and uncorrectable tracks of winners (those whom AI's decisions advantage) and losers (those disadvantaged). Losers find it difficult to escape that role as negative classifications feed data into AI that yields further negative classifications.

Should one use such a system as widely as possible to predict individual human action? It may seem obvious that one should not. AI creates winners and losers in ways that are mostly mistaken and uncorrectable, massively unfair, and as a prescription for social unrest, practically unwise. That does not settle the matter, however. After all, could not one justify using AI by showing that its benefits to society outweigh its costs? We understand benefits and costs broadly includes both quantifiable considerations and non-quantifiable ones (such as unfairness and social unrest). Many believe we can justify the extensive use of human predictive systems in just this way. For example,

> Because of the margin of uncertainty that edges all . . . [statistical] decisions, at least when honestly reached, we must collectively shoulder the burden of hope and fear [of being rightly or wrongly categorized], just as we must collectively submerge personal experience into public statistics and collectively stomach the possibility of local injustice in the name of global justice.[11]

Can AI can be justified in this way? If AI were not opaque, proponents of AI could try the following approach: show that AI takes all (or most) relevant costs and benefits into account, and then adequately justify the way in which the AI balances them. AI's opacity blocks this approach. Humans do not know what costs and benefits AI

[11] GERD GIGERENZER ET AL., THE EMPIRE OF CHANCE: HOW PROBABILITY CHANGED SCIENCE AND EVERYDAY LIFE 291 (1989).

considers, and its algorithm is opaque to humans. This leaves a single alternative: justifying AI by the consequences of the decisions it makes.

One, by no means small, problem is that AI makes no decisions before it is used, so the proponents cannot justify its initial use. It follows that even beginning to use AI is an offense to human freedom since fulfilling the *Knowledge Condition* requires a knowable justification for using AI. But, suppose for the sake of argument, humans do start using AI. Then this argument would be available; (1) in the past, the benefits of using AI outweigh the costs, and (2) it is likely that that they will continue to do so. It is *possible* for (1) to be true. Suppose AI allows us to cure diseases, to restore the climate, eliminate starvation, and order social relations in ways that yield a vibrant culture in which all have satisfying opportunities for self-realization. However, AI's results are likely to be much more mixed given that it creates winners and losers in mostly mistaken and uncorrectable ways. Massive unfairness and mistaken allocation of rewards and penalties is unlikely to generate net benefits. (2) is also problematic; a system's predictions are a function of the data it takes as input and the algorithm it employs.[12] Both will change as the aliens update the algorithm from time to time. Those changes can make AI's past decisions an uncertain guide to its future ones. The changes will be necessary because "predictive models tend to deteriorate over time — their ability to predict gets worse as economic, market and social change occurs. The relationships that were found between the predictor data and the outcome data when the model was originally constructed no longer apply."[13]

We assume there is no adequate justification for beginning to use or continuing to use AI. It follows that that the *Knowledge Condition* is not fulfilled, and that using AI is an offense to freedom.

B. Comparing Human-Created Systems

We should avoid using AI and sufficiently similar systems. This raises the following question: to what extent do human-created systems exhibit low accuracy, opacity, broad-based predictions, and feedback without error correction? The profile answers that question by identifying computational design features that meet three conditions:

---

[12] *See, e.g.*, JOHN D. KELLEHER & BRENDAN TIERNEY, DATA SCIENCE 143-144 (2018) ("Two major factors contribute to the [prediction] … that an ML [machine learning] algorithm will generate from a data set. The first is the data set the algorithm is run on . . . . The second factor … is the choice of ML algorithm.").

[13] STEVEN FINLAY, PREDICTIVE ANALYTICS, DATA MINING AND BIG DATA: MYTHS, MISCONCEPTIONS AND METHODS 79 (2014).

first, they are widely shared by current systems; second, they are reasonably accessible to legislators, governmental agencies, and researchers; third, they significantly affect the extent to which a system will exhibit low accuracy, opacity, broad-based predictions, and lack of error-correction. Our focus is broader than, but complementary to the focus on the use of classifications of race, gender, and sexual orientation.[14] As important as those examples are, they are also aspects of the more general problem we discuss.

II. PROFILING PREDICTIVE SYSTEMS

The profile consists of five computational design features: (1) level of accuracy; (2) how the data is collected and prepared; (3) choice classificatory and predictive targets; (4) the use of proxies to make classifications and predictions; (5) feedback — data feedback into the system and the presence or lack of error correction. This profile is sufficient for our purposes. For other purposes, one may want to alter or extend it. We consider these five computational design features in turn.

A. Level of Accuracy

The level of accuracy feature answers the question, "How accurate is the predictive system?"[15] Systems vary widely in accuracy.[16] Our concern, however, is with low accuracy systems predicting individual human action. This is why we stipulated that AI is highly inaccurate in a significant range of cases. The point was to make it similar to human-created systems that predict individual human action. Those systems are also highly inaccurate in a significant number of cases.[17] We offer two examples: direct mail advertising and the Chicago Police Department's Strategic Subject List algorithm.

In direct mail advertising, the task is to predict which consumers will respond positively (purchasing, signing up for a credit card, and so

---

[14] For greater focus on invidious discrimination, *see generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

[15] For our purposes it is sufficient to understand accuracy as the percentage of correct predictions. The use of accuracy in predictive analytics is more sophisticated and distinguishes among true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Then, where TOTAL = TP + FP + TN + FN, one can define accuracy as (TP + TN)/TOTAL and define the error rate as (FP + FN)/TOTAL. *See, e.g.*, KOTU & DESHPANDE, *supra* note 1, at 259.

[16] *See generally* HANNAH FRY, HELLO WORLD: BEING HUMAN IN THE AGE OF ALGORITHMS (2018).

[17] *See, e.g.*, FINLAY, *supra* note 13, at 6 ("[M]ost predictive models are quite poor at predicting how someone is going to behave"). *See generally*, O'NEIL, *supra* note 3 (discussing a number of cases of inaccurate predictions).

on) to an advertisement.[18] Humans are notoriously poor at identifying positive responders.[19] If a company, unaided by predictive analytics, mails an offer to a list of people, and with whom it has no prior relationship, about 1 percent of them will respond positively.[20] Using predictive analytics can improve the response rate by 20 to 30 percent.[21] That is still an error rate of 70 to 80 percent; however, in this case, improved predictive power can translate into increased profitability. If a direct mail marketer finds it profitable to mail with a 1 percent response rate, an improvement to 4 percent promises a significant increase in profitability.[22]

The Chicago Police Department's (CPD's) Strategic Subject List algorithm creates "a risk assessment score known as the Strategic Subject List or 'SSL.' These scores reflect an individual's probability of being involved in a shooting incident either as a victim or an offender."[23] The initial data consists information about arrests "contained within the CPD data warehouse."[24] Using that data, the algorithm constructs "social networks . . . to previous homicide victims to predict the likelihood of someone becoming a victim of a homicide."[25] The network is a "co-arrest" network. Chicago uses two types of co-arrests corresponding to two types of links in the network:

> A "first-degree" link refers to a relationship between a subject and an individual with whom the subject was previously co-

---

[18] *See* Hal Conick, *How to Use Direct Mail in the Modern Marketing Mix*, MARKETING NEWS, Sept. 2018, 14, 16.

[19] *See* KELLEHER & TIERNEY, *supra* note 12, at 153 ("Human intuition about customers can often miss important nonobvious segments or not provide the level of granularity that is required for nuanced marketing.").

[20] *See* FINLAY, *supra* note 13, at 7.

[21] *Id.* at 11.

[22] In general, the most favorable cases for the use of highly inaccurate predictive systems meet three conditions: (1) humans are even worse at prediction; (2) there is significantly increased benefit from improved prediction accuracy; (3) the costs of false positives and false negatives are low. The use of predictive analytics in direct mail marketing plausibly fulfils all three. Credit scoring is another example, a 20-30 percent improvement in credit scoring translates into granting 20-30 percent fewer loans to customers who would have defaulted or 20-30 percent more loans to good customers who will repay, depending upon how one decides to use the model. To put this in terms of raw bottom line benefit, if a bank writes off $500m in bad loans every year, then a reasonable expectation is that this could be reduced by at least $100m, if not more, by using predictive analytics. *See id.* at 2-6.

[23] *Strategic Subject List*, *supra* note 10. There is only one independent study of Chicago's SSL system, *see* Saunders et al., *supra* note 10, at 354 (evaluating the "pilot program developed in collaboration between the Chicago Police Department (CPD) and the Illinois Institute of Technology (IIT)."). Our discussion concerns that system, which has since undergone further development.

[24] Saunders et al., *supra* note 10, at 354.

[25] *Id.* at 354.

> arrested who later became a homicide victim. A "second-degree" link refers to a relationship in which a subject was co- arrested with another person who, in turn, was co-arrested with a later homicide victim.[26]

The underlying theory is that the more connections a person has to co-arrested individuals the more likely one will commit homicide or be a victim of one (depending on the nature of the links).[27]

The Strategic Subject List algorithm "identified less than 1% of homicide victims (3 out of 405)."[28] The case for using such systems is, to say the least, far less clear than the case for predictive analytics in direct mail advertising since the costs of focusing police attention on the wrong people can be very high. Unfortunately, "what happens when law enforcement agencies shift their analytic focus from street corners to people is unknown in the world of data-driven policing because there have been few formal evaluations . . ."[29]

Why is low accuracy a characteristic of systems that predict individual human action? The answer provides an important insight into the difficulties of using predictive analytics to predict individual action and also provides a transition to the next feature of the profile, data collection and preparation.

The fundamental reason is that the predictor data is decontextualized. An example illustrates what we mean by decontextualization; a social worker was tasked with the following:

> doing data entry for a contractor who was developing a tracking system for young people who were under state supervision. The frustration that finally drove her to quit the job was that the architecture of the database didn't allow social service workers to include narrative information about the context of kids' behavior. Simply, the system tracked each student's "success" or "failure" in a number of different programs. So, for example, if students stopped going to an afterschool program because they faced a serious crisis — a death in the family or an apartment fire, for example — a caseworker worker was forced to check a box that reported

---

[26] *Id.*

[27] *Id.*

[28] *Id.* at 363.

[29] ANDREW V. PAPACHRISTOS & MICHAEL SIERRA-ARÉVALO, OFFICE OF COMMUNITY POLICING SERVICES, POLICING THE CONNECTED WORLD: USING SOCIAL NETWORK ANALYSIS IN POLICE-COMMUNITY PARTNERSHIPS 22 (2018). However, the accuracy of the bail and sentencing algorithm is unclear. *See* Jordan Pearson, *Bail Algorithms Are as Accurate as Random People Doing an Online Survey*, MOTHERBOARD (Jan. 17, 2018), https://motherboard.vice.com/en_us/article/paqwmv/bail-algorithms-compas-recidivism-are-as-accurate-as-people-doing-online-survey.

> that they failed to complete the program. Because there was no input box for narrative case notes, there was literally no place in the system to account for the (sometimes pages of) contextual information written in the social workers' reports.[30]

The complaint is that the categories omit the contextual information necessary to understand and explain why the student acted as he or she did.[31] We explain human action through narratives that integrate values, purposes, intentions, and the context in which they occur into a meaningful pattern. One could, of course, add a checkbox for "death in the family" or "apartment fire," but that would still fail to capture the values, purposes, and intentions behind the student's reaction to those events. No set of checkboxes, however elaborate, will constitute a narrative integrating context, values, purposes, and intentions into a meaningful pattern.

Contextually rich narratives are the "data" on which human beings based their predictions and explanations of others' actions. As a further illustration, imagine two scenarios in which you are trying to predict whether Victoria will remain married to Victor once their children graduate from college. In the first, you know that Victoria's publicly observable behavior has been typical for a spouse in a twenty-year long first marriage, and you know that about 40 percent first marriages end in divorce. Can you predict *Victoria's* action? Will *she* divorce? The most the data allows you to do is make the statistical prediction that there is a forty percent chance of divorce. Compare knowing that Victoria regards her marriage as loveless, places a large disvalue on remaining in loveless relationships, and intends to divorce Victor when their children graduate from college. With the much context filled in, you can confidently predict that Victoria will divorce Victor.

Contrast the data for predictive analytics. As we explain in detail in the next subsection, it is decontextualized in the way the tracking system for supervised youth example illustrates. That is no accident. Decontextualization is inevitable in statistical explanation. To make statistical predictions about people, you look for regularities that hold with some degree of probability for people in certain categories. To achieve this goal, one abstracts from the enormous variation in individuals' life histories and looks for reliable correlations between

---

[30] VIRGINIA EUBANKS, DIGITAL DEAD END: FIGHTING FOR SOCIAL JUSTICE IN THE INFORMATION AGE 95 (2011).

[31] *See generally* Machiel Keestra, *Understanding Human Action: Integrating Meanings, Mechanisms, Causes, and Contexts*, *in* CASE STUDIES IN INTERDISCIPLINARY RESEARCH 225 (Allen F. Repko, William H. Newell & Rick Szostak eds., 2011).

categories that are independent of the idiosyncratic paths people traced to get into those categories.[32] It is no surprise then that predictive systems predicting individual human action are inaccurate.

Some may object that we have missed the point. Is one of the key points of predictive analytics that the computer's ability to sift through massive amounts of data enables predictive analytics to do what humans cannot? Will a massive amount of decontextualized data allow a system to predict individual action? Not yet, at least. Current predictive analytics cannot make reasonably accurate predictions about human action by extracting information about values, purposes, and intentions from decontextualized data and then reasoning about what behavior that information predicts. Professor Barbara J. Grosz observes about currently popularly deep-learning[33] approaches to predictive analytics:

> [T]hese systems are really good at statistical learning, pattern recognition and large-scale data analysis, but they don't go below the surface. They can't reason about the purposes behind what someone says. Put another way, they ignore the intentional structure component of dialogue. Deep-learning based systems more generally lack other hallmarks of intelligence: they cannot do counterfactual reasoning or common-sense reasoning.[34]

Data decontextualization is a key factor in the explaining the low accuracy of systems predicting human action. Data decontextualization

---

[32] *See* GIGERENZER ET AL., *supra* note 11, at 184 ("In social science, as opposed to molecular physics, it is possible to trace individual life histories and, as we may (counterfactually) grant for the sake of the argument, explain them in terms of determining causal chains. But for the sociological purpose of explaining the overall structure of a society and its changes, this 'historical' or 'dynamical' treatment . . . would have to give way to a structural treatment, one form of which is statistics. In order for a statistical treatment to make sense, the overall structures must be invariant with respect to changes in the many detailed histories (of molecules or of people).").

[33] *See* FINLAY, *supra* note 1, at 128-29 ("'Deep' neural networks (Deep learning/Deep belief networks) are very large and complex neural networks (often containing thousands or millions of artificial neurons) which are used for 'AI' tasks such as speech recognition and in self-driving cars. . . . [Neurons are the key component] of a neural network . . . . [A] neuron is a linear model whose score is then subject to a (non-linear) transformation. A neural network can therefore be considered as a set of interconnected linear models and non-linear transformations.")

[34] MARTIN FORD, ARCHITECTS OF INTELLIGENCE: THE TRUTH ABOUT AI FROM THE PEOPLE BUILDING IT 338 (2018). The mathematician Hannah Fry makes a similar point:

> Although AI has come on in leaps and bounds of late, it is still only 'intelligent' in the narrowest sense of the word. It would probably be more useful to think of what we've been through as a revolution in *computational statistics* than a revolution in intelligence. . . . [That is] a far more accurate description of how things currently stand.

FRY, *supra* note 16, at 12.

occurs during data collection and preparation, which is the next feature in the profile.

B. Data Collection and Preparation

The "data preparation" component of the profile answers the following question: what data is collected in the first place, and then what is kept and what is eliminated? The answer matters: "[e]verything that isn't counted as relevant is then marginalized and rendered invisible to our models."[35] Data collection and preparation renders contextual information invisible in five ways: collection, selection, cleaning, structuring, and choice of attributes. The processes typically overlap and interact,[36] but, for convenience, we treat them as separate and distinct. Those processes create the decontextualized "reality" which a predictive system uses to make its predictions.[37] The elimination of data involved in collection and selection is significant for three reasons. First, it contributes to inaccuracy by decontextualizing the information. Second, as we explain below when discussing opacity, it can lead to the failure of the *Knowledge Condition*. Third, it can create objectionable bias in the predictive system, as we explain when discussing the classificatory and predictive targets part of the profile.

1. Collection and Selection

Data preparation begins with data collection, a starting point that is already a stream of decontextualized data. Much of the information in databases comes from the data detritus people leave behind. Those leftovers are hardly precise indicators of the contexts of their creation. To see why, note that the data typically divides into meaningful content and data about the content (metadata, information about the time of creation or transmission, device used, and so on).

The content does not fully (or often even significantly) indicate the relevant context. When communicating content, people *assume* that they and their audiences understand a background that includes the relevant context, values, purposes, and intentions. They do not make that background explicit in the communication, and even a full record of the words and images exchanged would not capture it.

It may seem that metadata, however, captures a great deal of contextual information. Email metadata, for example, can include

---

[35] CATHY O'NEIL, ON BEING A DATA SKEPTIC loc. 99 (2013) (ebook).

[36] *See* KELLEHER & TIERNEY, *supra* note 12, at 58-60.

[37] *Id.* at 66 ("With regard to getting the right data for a project, a survey of data scientists in 2016 found that 79 percent of their time is spent on data preparation.").

sender's name, email and IP address, recipient's name and email address, plus a great deal more.[38] It is indeed true that aggregating metadata can reveal a great deal about people,[39] but it typically does not reveal an explanatory pattern of values, purposes, and intentions. Consider this search metadata for example:

> [2018/03/09 18:34:44] abortionfacts.com
> [2018/03/09 18:35:23] plannedparenthood.org
> [2018/03/09 18:42:29] dcabortionfund.org
> [2018/03/09 19:02:12] maps.google.com

The data reveals a user's concern with abortion, but it does not reveal why. The searcher could be a woman seeking an abortion, a pro-abortion activist, an anti-abortion activist, or an academic researcher. You can eliminate some of these possibilities by adding more data (for example, that the searcher is male), but no compilation of data, however extensive, will constitute an integrated narrative revealing values, purposes, and intentions. As Shoshana Zuboff notes, data collection practices

> render the entire world's actions and conditions as behavioral flows. Each rendered bit is liberated from its life in the social, no longer inconveniently encumbered by moral reasoning, politics, social norms, rights, values, relationships, feelings, contexts, and situations. In the flatness of this flow, data are data, and behavior is behavior.[40]

In creating a predictive system, the next step after data collection is data selection. It is rare for all the collected data to be used. Instead, one picks and chooses. As the data scientist Steven Finlay explains:

> One feature of Big Data is that most of it has a very low information density, making it very difficult to extract useful customer [or other] insights from it. A huge proportion of the Big Data out there is absolutely useless when it comes to forecasting consumer [or other] behavior. You have to work pretty hard at finding the useful bits that will improve the

---

[38] Rebecca Greenfield, *What Your Email Metadata Told the NSA about You*, THE ATLANTIC (June 27, 2013), https://www.theatlantic.com/technology/archive/2013/06/email-metadata-nsa/313842/.

[39] *See generally* Jonathan Mayer, Patricia Mutchler & John C. Mitchell, *Evaluating the Privacy Properties of Telephone Metadata*, 113 PROC. NAT'L ACAD. SCI. 5536 (2016).

[40] SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 211-12 (2019).

accuracy of your predictive models . . . [41]

One selects data that will serve one's predictive goals. One is looking for reliable correlations between categories that are, as we put it earlier, independent of the idiosyncratic paths people traced to get into those categories. Thus, to the extent that collected data is not relevant to the predictive task at hand, one does not include it in the predictive system. This further liberates data "from its life in the social, no longer inconveniently encumbered by moral reasoning, politics, social norms, rights, values, relationships, feelings, contexts, and situations."[42]

2. Cleaning and Structuring

Cleaning and structuring are processes through which selected data gets organized, altered, and placed into a database. They are key processes that help construct the database "reality." In discussing cleaning and structuring, we will use this simple example of a traditional relational database,[43] organized as a table of rows and columns:

Owner	**Make**	**Model**	**Occupation**	**City**	**Age**
John Smith	Mazda	Miata	Teacher	Chicago	44
Joe Friday	Ford	Fairlane	Police Officer	Los Angeles	39

Each row contains a data set referring to a single item or "record," in this case car owners and their car. Each column consists of one "attribute" of an item. In this case, 'Owner', 'Make', 'Model', 'Occupation', and 'City' are all "attributes" of the records. Entries indicate the values of attributes; for example, Joe Friday owns a Ford Fairlane, is a police officer living in Los Angeles, and is thirty-nine years old. Note that nothing in the database reveals the contextually rich narrative that explains why, at age thirty-nine, Friday became a Police Officer, owns a Ford Fairlane, and lives in Los Angeles.

---

[41] FINLAY, *supra* note 13, at 14. *See also id.* at 161 ("Whatever data you have, wherever it has come from, whether it's structured or unstructured, data in its raw form is not often very useful for prediction. Usually, there is far greater merit in deriving new types of data from it rather than using it as it is.").

[42] ZUBOFF, *supra* note 40, at 211.

[43] For a discussion of databases, *see generally* ABRAHAM SILBERSCHATZ, HENRY F. KORTH & S. SUDARSHAN, DATABASE SYSTEM CONCEPTS (6th ed. 2010).

Structuring puts data in the proper form to include in a database. As Finlay notes, building a predictive system requires structuring the data appropriately:

> All of the methods used to create predictive models require data to be well structured, and the data must be categorical (e.g. occupation, marital status and gender) or numeric (e.g. age, income and time at address). A predictive model can't be built if the data is not in one of these two formats.[44]

Getting the data in the proper form typically requires cleaning it:

> Data is dirty, filthy, messy stuff. Often it's incorrect, missing or badly formatted, particularly where humans have been involved in creating and/or collecting it. Sometimes numeric data is held as text, or text data is forced into fixed-length fields resulting in some data being truncated, and so on. Consequently, a lot of the time and effort . . . can be spent "cleaning" the data before it's ready to be used.[45]

Cleaning and structuring the data may eliminate or alter information when properly formatting it, interpreting truncated data, fitting data into fixed length fields, discarding records that have some missing attributes or making up values for the missing attributes, and in correcting or discarding information seen as incorrect. To take just one example, "[i]f unstructured data such as text or images have been considered for inclusion, then suitable text/image analytics will have been applied to extract the useful bits, so as to create a suitable structured representation of that data."[46]

To summarize, cleaning and structuring data removes context and frequently also alters or removes some data.

3. Choosing Attributes

Choosing attributes is the process of determining the labels for the top of the database. In our earlier example, the attributes are Owner, Make, Model, Occupation, City, and Age:

Owner	**Make**	**Model**	**Occupation**	**City**	**Age**
John Smith	Mazda	Miata	Teacher	Chicago	44
Joe Friday	Ford	Fairlane	Police Officer	Los Angeles	39

[44] Finlay, *supra* note 13, at 177.

[45] *Id.* at 160.

[46] Finlay, *supra* note 13, at 165.

In the case of predictive analytics, a predictive goal guides the choice of attributes. "The role of the data analyst is to create *informative* features: those would allow the learning algorithm to build a model that predicts well."[47] Suppose you want to predict when current car owners will buy another car. You have a wealth of data about current car owners, including their names of owners, the makes and models of their current cars, their ages, occupations, and cities of residence. To the extent you have reason to think that that information plays a significant role in predicting when an owner will buy another car, you have reason to pick Owner, Make, Model, Occupation, City, and Age as attributes.

Our deliberately simple example involves just six attributes. But, as Finlay notes (talking about variables, another term for attributes):

> For many projects there are potentially millions of . . . variables that could be considered — far more than any analytical system can deal with. Therefore, an important part of the model-building phase is using business [or other relevant] knowledge to come up with ideas as to what types of . . . variables one should consider.[48]

Choosing variables and filling in their values for each identifier in the database constructs the decontextualized database "reality" that will serve as the data input to the systems predictions.

Data preparation is an important source of opacity (which, as the aliens' system AI illustrates, can be a bar to fulfilling the *Knowledge Condition*).

4. Opacity

We define opacity as follows; a predictive system is *opaque* to the extent one cannot identify the factors that determine its predictions and explain (in a human-understandable way) how those factors yield the predictions. On a narrow view, just two factors determine a predictive system's classifications and predictions: the database and the algorithm run on that database.[49] We understand opacity more broadly to include the preparation of the database as a factor. The rationale is that what is omitted from or altered in database may be an important factor in explaining a system's predictions, as the following example illustrates.

---

[47] ANDRIY BURKOV, THE HUNDRED-PAGE MACHINE LEARNING BOOK 43-44 (2019) (emphasis in original).

[48] Finlay, *supra* note 13, at 162.

[49] *See, e.g.*, KELLEHER & TIERNEY, *supra* note 12, at 143-144. ("Two major factors contribute to the [prediction] … that an ML [machine learning] algorithm will generate from a data set. The first is the data set the algorithm is run on . . . . The second factor … is the choice of ML algorithm.").

Suppose Sally defaults on a $50,000 credit card debt. When the credit card company begins collection procedures, she declares bankruptcy. Roger also defaults on a $50,000 credit card debt and declares bankruptcy. Sally incurred her debt to pay for lifesaving treatment for her eight-year old daughter. Roger incurred his debt through compulsive gambling, an addiction which has resisted years of attempted cures. Suppose that post-bankruptcy, Sally is a good credit risk. Her daughter is well with no further expenses expected, and she is earning a good income that considerably exceeds her expenses. Roger, however, remains a poor risk. Imagine a credit-scoring predictive system that predicts that both Sally and Roger are similar risks because of their bankruptcies. There are two explanations. One appeals to just the database and the algorithm. Given the bankruptcy information in the database, Sally and Roger look roughly the same to the algorithm which then classifies them similarly. The second explanation adds an additional fact to the first, a fact that was omitted from the system's database; Sally's bankruptcy was the result of a medical emergency. The system sees Sally and Roger as similar risks because it ignores the different contextual explanations of why Sally and Roger went bankrupt. The second explanation is relevant to regulating predictive systems to the extent that one wants to impose requirements on what databases must, may, and must not contain.

Unfortunately, the information needed for explanations of the second sort is rarely available and hence predictive systems are in this way typically opaque. Businesses routinely guard both the database (as well as the predictive system itself) as trade secrets, and even when the database is available, you typically do not know what data was discarded or altered in the processes of collection, selection, cleaning, and structuring. Our focus on database opacity may surprise some. Discussions of the opacity typically focus on the algorithm, not the database.[50] Those discussions see the algorithm's unavailability and the complexity of its source code as the cause of opacity. It is unavailable because businesses typically guard it as a trade secret. Even when it is available, it is typically so complex that "[t]he source code of computer systems is illegible to nonexperts. In fact, even experts often struggle to understand what software code will do: inspecting source code is a very limited way of predicting how a computer

[50] *See e.g.*, PASQUALE, *supra* note 14. Pasquale does contend that "without access to the underlying data and code, we will never know what type of tracking is occurring, and how the discrimination problems long documented in 'real life' may even now be insinuating themselves into cyberspace." *Id.* at 40. But his main concern is with source code, and it is not clear that he has in mind the *preparation* of the database as distinct from the end result of the database itself.

program will behave."[51] We do not disagree. Our point is that databases should join source code as a cause of opacity.

C. Classificatory and Predictive Targets

One typically designs a predictive system in order to make one or more types of classifications or predictions. The "classificatory and predictive targets" part of the profile answers the question, what and how does the system classify and predict? To discuss this part of the profile, we first discuss three concepts — explainability, opacity, and transparency — and then explain their potential regulatory relevance.

1. Explainability, opacity, and transparency

In the computer science literature, the availability of an answer to the question is an issue of explainability.[52] A predictive system is *explainable* if one can provide an adequate, human-understandable characterization and explanation of its classifications and predictions, where explanation appeals to just two factors: the particular database involved, and the particular algorithm used. Some predictive systems are typically explainable; others are not. Decision trees are an example of an explainable algorithm.[53] One can readily characterize the output. For example, "our system will predict that the 20-something will go to the movies if their parents are visiting, otherwise play tennis if it's sunny and not windy, go to the movies if it's sunny, windy, and they don't have much money . . . " Once a prediction is made, a human-understandable explanation of why that prediction was made for that particular predictor data is typically readily available. On the other hand, deep neural nets and support vector machines, for example, are sufficiently complex that they usually are not explainable.

A system is *transparent* if it is not at all opaque; so, opaque versus transparent are sliding scale opposites. Full transparency requires explainability — a human-understandable explanation of why the system generates its classifications and predictions — but explainability is not sufficient. A system may be both explainable and opaque. The decision tree system we imagined above, for example, is

---

[51] Kroll et al., *supra* note 3, at 638.

[52] *See generally* Mark G. Core et al., *Building Explainable Artificial Intelligence Systems*, *in* PROC. OF THE TWENTY-FIRST NAT'L CONF. ON ARTIFICIAL INTELLIGENCE 1766 (2006); David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence (last visited Apr. 6, 2019); Wojciech Samek, Thomas Wiegand & Klaus-Robert Müller, *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*, ARXIV (Aug. 28, 2017), https://arxiv.org/pdf/1708.08296.pdf.

[53] *See, e.g.*, Finlay, *supra* note 1, at 44.

explainable but would be opaque if relevant details about the construction of its database were unavailable.

Explainability and opacity versus transparency are relevant to regulation. One may want to require explainability but not transparency in some cases (for example, to allow the use of opaque but explainable decisions trees in some case) and in other cases one may want to limit opacity by imposing requirements on the construction of the database.

One might well be concerned with both explainability and broader issues of transparency in responding to predictive systems that generate a numerical ranking that purports to predict a consumer's value to a business.[54] As eBureau, one of the companies offering such scoring, explained, "eBureau's patented technology analyzes vast amounts of predictive data to help you with critical decisions throughout the customer lifecycle."[55] Here is how eBureau (now a part of Transunion) worked:

> A client submits a data set containing names of tens of thousands of sales leads it has already bought, along with the names of leads who went on to become customers. EBureau then adds several thousand details — like age, income, occupation, property value, length of residence and retail history — from its databases to each customer profile. From those raw data points, the system extrapolates up to 50,000 additional variables per person. Then it scours all that data for the rare common factors among the existing customer base. The resulting algorithm scores prospective customers based on their resemblance to previous customers.[56]

Businesses use e-scores to determine how to treat consumers in a variety of situations:

> A growing number of companies, including banks, credit and debit card providers, insurers and online educational institutions are using these scores to choose whom to woo on the Web. These scores can determine whether someone is pitched a platinum credit card or a plain one, a full-service cable plan or none at all. They can determine whether a customer is routed promptly to an attentive service agent or

---

[54] *See, e.g.*, Natasha Singer, *Secret E-Scores Chart Consumer Buying Power*, N.Y. TIMES (Aug. 18, 2012), https://www.nytimes.com/2012/08/19/business/electronic-scores-rank-consumers-by-potential-value.html. Numerical rankings as output are not confined to e-scores. In general, a "model's predictions are almost always represented by a single number — a *score*." STEVEN FINLAY, PREDICTIVE ANALYTICS IN 56 MINUTES 4 (2015) (emphasis in original).

[55] *About Us*, EBUREAU, http://www.ebureau.com/about [https://web.archive.org/web/20171016113624/http://www.ebureau.com/about] (last visited Apr. 6, 2019).

[56] Singer, *supra* note 54.

relegated to an overflow call center.[57]

The scoring systems raise concerns about both opacity and explainability. Concerns about opacity arise because information about data preparation is unlikely to be available since the will be almost certainly be guarded as a trade secret. Concerns about explainability arise because the scoring systems frequently use classification algorithms with poor explainability. For example, they may use clustering to sort consumers into "similarly behaving" groups.[58] A "significant weakness of clustering [models] . . . is their complexity and 'black box' nature. You can't tell by looking at these types of model what variables contributed significantly to the model score and which did not."[59] eBureau's system, for example, is likely complicated enough that a detailed picture of what it predicts is difficult if not impossible to obtain. "With eScores' automated statistical modeling software, over 25,000 variables are commonly incorporated in the model development process, generating superior score performance. eBureau's highly scalable system allows the number of modeling attributes to grow as eBureau's data resources expand." [60]

2. Training data bias

Choices of classificatory and predictive targets can have objectionably discriminatory results. We consider two examples. One concerns the general use of training data in the machine learning approach known as supervised learning that is used in developing all or almost all predictive analytics systems.[61] The second example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a bail and sentencing algorithm that, incidentally, was created using supervised learning, but that is not the point we emphasize here.[62] COMPAS also illustrates a problem inherent in its underlying data.

---

[57] *Id.*

[58] *See* O'NEIL, *supra* note 3, at 145 ("[E-scores] carry out thousands of 'people like you' calculations. And if enough of these 'similar' people turn out to be deadbeats or, worse, criminals, that individual will be treated accordingly.").

[59] Finlay, *supra* note 13, at 124.

[60] *eScores Data Sheet*, EBUREAU (2010), http://www.ebureau.com/sites/all/files/file/datasheets/ebureau_escore_datasheet.pdf [https://web.archive.org/web/20160208005758/http://www.ebureau.com/sites/all/files/file/datasheets/ebureau_escore_datasheet.pdf].

[61] *See* FORD, *supra* note 35, at 18.

[62] *See* Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, *in* PROC. OF THE THIRTIETH CONF. ON NEURAL INFO. PROCESSING SYS. 3323 (2016) (discussing criteria for fairness that arise in supervised learning systems such as COMPAS).

Supervised learning requires that you have some data available that includes the outcome data you want to classify or predict. For example, you want to predict who will default on credit card debt and you have data from some past cases, showing the types of people who did and did not default on credit card debt. The overall goal is to build a predictive system that will give good answers on new data where the outcome data is *not* available to you, for example, new credit card applicants. In supervised learning, you collect data — the training data — and choose a type of classification or prediction algorithm (e.g., clustering, decision trees, or neural nets) for your predictive system. Then, corresponding to each type of classification or prediction algorithm is one or more *training algorithms* that convert the training data into a classification or prediction algorithm.[63] As other scholars have shown, bias in the training data will translate into bias in the predictions made when the algorithm is in use.[64]

A classic example is St. George's Hospital Medical School in London.[65] In the 1970's, it developed a computer program to "cull down the two-thousand applications to five-hundred, at which point humans would take over."[66] They trained the program on the years of human-rated applications it had in its records. The program developed its criteria from the training data and thereby incorporated the human biases reflected in that data.[67] "In 1988, the British government's Commission for Racial Equality found the medical school guilty of racial and gender discrimination in its admissions policy."[68]

3. Another Form of Bias in the Data

The judicial bail and sentencing algorithm COMPAS illustrate another source of bias.[69] COMPAS predicts the likelihood that a person convicted of a crime will commit another in the future. Of course, it is extremely undesirable for a system like COMPAS to exhibit racial bias. Thus, one would hope COMPAS would have the following two

---

[63] Typically, one runs a training algorithm on part of the training data and sees how well the resulting prediction algorithm predicts outcome data for the remaining training data. If the predictions are not as accurate as desired, one can repeatedly make various adjustments or try other training algorithms. *See generally* TOM M. MITCHELL, MACHINE LEARNING (1st ed. 1997); BURKOV, *supra* note 47.

[64] O'NEIL, *supra* note 3, at 115-118.

[65] *Id.* at 115-116.

[66] *Id.*

[67] *Id.*

[68] *Id.* at 117.

[69] Jeff Larson & Julia Angwin, *Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*, PROPUBLICA (Dec. 30, 2016), https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

features: first, COMPAS makes equally accurate predictions regardless of race; second, it makes false positive and false negative mistakes at the same rate for all racial groups.[70] The problem is that it is mathematically impossible to meet both these conditions if the fractions of people who commit crimes differ for racial groups.[71] This is currently the situation in the United States where African Americans commit more crimes than whites (a result of centuries of systematic discrimination).[72] The consequences of the conditions the COMPAS system implemented? "[B]lack defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk."[73]

4. Infeasible Classifications and Predictions

The classifications and predictions a system is designed to make should be feasible. Current teacher rating systems fail to meet this requirement. Those systems use "value-added" as the predictive target. It is a simplification, but still essentially correct, to characterize the systems as measuring value-added by measuring the performance of students on two standardized tests, one at the beginning of instruction, the other at the end.[74] The difference in scores is the "value-added." It is clear that value-added systems fail to distinguish good from bad teachers — often classifying bad as good and good as bad.[75] There are,

---

[70] FRY, *supra* note 16, at 61-63.

[71] *Id.*

[72] *Id.* at 66-69.

[73] Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[74] *See* Steven Glazerman et al., *Evaluating Teachers: The Important Role of Value-Added*, BROOKINGS (Nov. 17, 2010), https://www.brookings.edu/research/evaluating-teachers-the-important-role-of-value-added/ ("The latest generation of teacher evaluation systems seeks to incorporate information on the value-added by individual teachers to the achievement of their students. The teacher's contribution can be estimated in a variety of ways, but typically entails some variant of subtracting the achievement test score of a teacher's students at the beginning of the year from their score at the end of the year, and making statistical adjustments to account for differences in student learning that might result from student background or school-wide factors outside the teacher's control. These adjusted *gains* in student achievement are compared across teachers. Value-added scores can be expressed in a number of ways. One that is easy to grasp is a percentile score that indicates where a given teacher stands relative to other teachers. Thus a teacher who scored at the 75th percentile on value-added for mathematics achievement would have produced greater gains for her students than the gains produced by 75 percent of the other teachers being evaluated.").

[75] *See* Gary Rubinstein, *Analyzing Released NYC Value-Added Data Part 2*, GARY RUBINSTEIN'S BLOG (Feb. 28, 2012), https://garyrubinstein.wordpress.com/2012/02/28/analyzing-released-nyc-value-added-data-part-2/; O'NEIL, *supra* note 3, at 134-140.

nonetheless, "a slew of proprietary models, being sold for the most part by private education consulting companies, that purport to measure the 'value added' by a given teacher through the testing results of their students from year to year."[76]

D. Proxies

One uses proxies when one cannot directly measure the things relevant to the predictions one would like to make. Suppose, for example, a teacher is interested in the length of time students pay attention in class. She cannot directly measure paying attention, so she uses proxies: taking notes, looking at material displayed on the board, and so on. The "proxy" dimension of a system's profile is determined by the answer to the question, "What are the proxies, and how wide reaching are the predictions they support?"

1. Broad-based Predictions

Auto insurance illustrates broad-based predictions. As a *Consumer Reports* study of rates notes, "behind the rate quotes is a pricing process that judges you less on driving habits and increasingly on socioeconomic factors. These include your credit history, whether you use department-store or bank credit cards, and even your TV provider."[77] For example, "[i]n New York State . . . a dip in a driver's credit rating from 'excellent' to merely 'good' could jack up the annual cost of insurance by $255."[78] Grant, for the sake of argument which may well not be true,[79] that there is some statistical correlation between credit scores and driving safety. A normative question remains: *should* a drop in your credit score raise your car insurance premium no matter what the reason for the drop? A variety of scenarios can lead to a reduced credit rating, and in some one may think the debtor's actions praiseworthy — for example, incurring debt to pay for a child's education or health care. Especially in those cases, why should the events in one area of one's life penalize one in another? The more predictive systems use proxies that reach across a variety of areas of a person's life, the more a negative classification by one system can lead to a negative classification by another. This can make it quite difficult

---

[76] O'NEIL, *supra* note 35, at loc. 115.

[77] *Special Report: Car Insurance Secrets*, CONSUMER REPORTS (July 30, 2015), https://www.consumerreports.org/cro/car-insurance/auto-insurance-special-report/index.htm.

[78] O'NEIL, *supra* note 3, at 164.

[79] Eric Zorn, *What Does My Credit Rating Have to Do with My Driving?*, CHI. TRIBUNE (Aug. 6, 2015), http://www.chicagotribune.com/news/opinion/zorn/ct-auto-insurance-credit-rating-perspec-zorn-0807-20150806-column.html.

to recover from financial or personal difficulties as lower ratings in some areas generate lower ratings in others. As one commentator notes, "as more of our life is quantified . . . proxy judgments can get more esoteric yet more intrusive. Better prediction can lead to subtler and more nefarious discrimination."[80]

2. Explainability

The use of proxies can create explainability issues. Recall that eBureau's "automated statistical modeling software . . . allows the number of modeling attributes to grow as eBureau's data resources expand."[81] If the automated growth creates proxies, there is no guarantee that eBureau will know what they are. Without that knowledge, they may be unable to provide a human-understandable explanation of how the system reaches its classifications and predictions.

E. Feedback Mechanisms

The "feedback" dimension of the profile asks different two questions, "Does the system have a mechanism for error correction?", and "Do the classifications and predictions of the system have consequences that create information that feedback into the data on which the system bases future classifications and predictions?"

1. Error Correction

Without error-correcting feedback, a predictive system "can continue spinning out faulty and damaging analysis while never learning from its mistakes."[82] Suppose, for example, that Amazon's predictive systems "started recommending lawn care books [primarily] to teenage girls, the clicks would plummet . . . "[83] Monitoring click rates, however, provides error correcting feedback, and "the algorithm would be tweaked until it got it right."[84]

Unfortunately, many human predictive systems lack error-correcting feedback. E-scores are a good example. Suppose a business uses e-scores to decide which callers to route to a person, and which callers to route to a series of voice prompts. It is not likely to have a way to determine it consigned a potentially valuable customer to an

---

[80] SETH STEPHENS-DAVIDOWITZ, EVERYBODY LIES: BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE 262 (2017).

[81] *eScores Data Sheet*, *supra* note 60.

[82] O'NEIL, *supra* note 3, at 7.

[83] *Id.*

[84] *Id*.

infuriating series of voice prompts. Even if it did, it is unlikely to have a way to enter that information in an error correcting way into the e-score system. Additional examples of lack of error-correcting feedback include teacher rating systems,[85] judicial sentencing algorithms,[86] predictive policing systems,[87] *US News and World Report* rankings of colleges and universities,[88] and pre-hiring screening systems.[89]

2. Feedback into Predictor Data

Now we turn to the question of whether the classifications and predictions of the system have consequences that create information that feedbacks to create additional predictor data. This is typical of "[m]achine-learned models for recommendation, ranking, and spam detection [that] all become obsolete unless they are regularly infused with new data."[90] Data feedback, especially when combined with lack of error correction and broad-based predictions, creates mistaken and uncorrectable tracks of winners and losers.

F. Avoiding Human-Created Alien Intelligence

Can we regulate the features in the profile in a way that fulfills the *Knowledge Condition*: those subject to predictive systems are able with reasonable effort to know that there is an adequate justification for the decisions? Further, can we do so in ways that ensure that human-created systems do not exhibit the features that make AI objectionable? Our answer to both questions assigns a key regulatory role to the Federal Trade Commission (FTC). We explain how to use FTC actions to create norms. The norms create market incentives for businesses to use predictive systems with profiles that minimize the objectionable features of AI and also ensure that the *Knowledge Condition* is fulfilled.

We begin with the *Knowledge Condition*.

III. FULFILLING THE *KNOWLEDGE CONDITION*

Respecting human freedom requires fulfilling, or adequately approximating, the *Knowledge Condition*. This requirement may seem impossible to fulfill. When a 25,000 variable, constantly updated e-score system shunts a consumer into a long series of voice mail

---

[85] *Id.*

[86] *Id.* at 25-27.

[87] *Id.* at 86-89.

[88] *Id.* at 53-54.

[89] *Id.* at 110.

[90] Alekh Agarwal et al., *Making Contextual Decisions with Low Technical Debt*, ARXIV (last updated May 9, 2017), http://arxiv.org/abs/1606.03966.

prompts, how is he or she supposed to determine whether there is an adequate justification for his or her treatment? In this section, we explain in general how norms can provide an answer to such questions. In the next section, we sketch a regulatory proposal that leads to norms that provide an answer to that question in the case of predictive systems.

A. Coordination Norms

The norms that concern us are *coordination norms*, a subspecies of norms generally. Family holiday dinners illustrate coordination norms. Imagine a congenial family, all of whose members share the goal of a harmonious dinner. As everyone realizes, this requires a selective flow of information. There are things one can tell Aunt Jane that must not reach Uncle John's ears, and so on. The family members know and observe the required strictures. The example illustrates three conditions that, when generalized, define coordination norms: (1) There is a behavioral regularity: family members collectively ensure the desired selective flow of information; (2) They adhere to that regularity to achieve a shared goal: harmonious relations; (3) They conform only as long as enough other members do so. Only collective conformity can ensure harmony. So, unless enough family members conform, there is little point in any one member's conforming. In general, a coordination norm is a behavioral regularity in a group, where the regularity exists at least in part because everyone thinks that, in order to realize a shared goal, he or she ought to conform to the regularity as long as everyone else does.

For another example, imagine you are about to enter an elevator with two people already in it. They are standing near the opposite walls roughly in line with each other. Where do you stand? Behind them and equidistant from them. Why? Because that is the norm. The elevator norm is to stand as far away as you can from the person nearest to you. More fully, the behavioral regularity is that elevator users maximize the distance from the person nearest them to realize the shared goal of using elevators while minimizing the sense of overcrowding. The regularity exists because people think they ought to conform to the norm to realize the goal as long as they trust others do. If others just stand where they like, being a unilateral distance maximizer is pointless; it does not prevent overcrowding.

We are concerned with a subclass of coordination norms — those that are also informational norms. *Informational norms* are social norms that constrain the collection, use, and distribution of

information.[91] The family holiday dinner illustrates a coordination norm that is also an informational norm. In general,

> [Informational] norms circumscribe the type or nature of information about various individuals that, within a given context, is allowable, expected, or even demanded to be revealed. In medical contexts, it is appropriate to share details of our physical condition or, more specifically, the patient shares information about his or her physical condition with the physician but not vice versa; among friends we may pour over romantic entanglements (our own and those of others); to the bank or our creditors, we reveal financial information; with our professors, we discuss our own grades; at work, it is appropriate to discuss work-related goals and the details and quality of performance.[92]

As Nissenbaum's examples illustrate, the contextual constraints on information flows vary with the social roles of the actors. The constraints are, as we will say, *role-appropriate*. Role-appropriate constraints create selective flows of information, different selective flows for different roles. The selectivity implements a tradeoff; it secures the benefits of information processing to an extent and protects privacy to an extent.

Not all information norms are also coordination norms, but the ones that concern us are. We offer two examples. To avoid misunderstanding, we should note that we describe the examples as if governmental and private sector surveillance were not a significant factor (a more or less mid-Twentieth Century description). This simplifying assumption is legitimate since the goal of this section to show how it is possible for norms to facilitate the fulfillment of the *Knowledge Condition*. The next section sketches a procedure for making that possibility a reality.

The first example concerns restaurants and their customers. There is a behavioral regularity; restaurants process customer information only in role-appropriate ways, where it is role-appropriate for the restaurant to collect, use, and distribute customers' personal information to meet the customers' restaurant needs. Further, restaurants and customers conform to the regularity because think they

---

[91] We have discussed coordination norms at length elsewhere. *See., e.g.*, SLOAN & WARNER, *supra* note 9; Robert H. Sloan & Richard Warner, *The Self, the Stasi, and the NSA: Privacy, Knowledge, and Complicity in the Surveillance State*, 17 MINN. J. L. SCI. TECH. 347 (2016); ROBERT H. SLOAN & RICHARD WARNER, UNAUTHORIZED ACCESS: THE CRISIS IN ONLINE PRIVACY AND INFORMATION SECURITY (2014).

[92] Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 138 (2004).

should in order realize the shared goal of meeting customers' restaurant needs as long as they trust each other to do so.

For the second example, suppose Victoria is visiting a brick-and-mortar bookstore. Here, there is a behavioral regularity; bookstores process customer information only in role-appropriate ways. It is role-appropriate for the bookstore to collect, use, and distribute customers' personal information to meet the customers' bookstore needs. Furthermore, bookstores and customers participate in role-appropriate information processing to realize the shared goal of meeting customers' bookstore needs. They do so as long as they trust others to do so.

B. Fulfilling the Knowledge Condition

Victoria in the bookstore illustrates how informational coordination norms facilitate the fulfillment of the *Knowledge Condition*. She knows the bookstore will process some range of personal information. The *Knowledge Condition* is fulfilled if she is able with reasonable effort to know that there is an adequate justification for the information processing. To see that this condition is fulfilled, first ask, what exactly does Victoria want to know? What would count for her as an adequate justification? We assume that Victoria and consumers generally want to know if the tradeoff between benefits and costs is acceptable — where "benefits" and "costs" include both quantitative and non-quantitative considerations. Victoria knows that the tradeoff is acceptable provided she knows two things; her transaction with the bookstore is governed by an appropriate informational coordination norm, and the tradeoffs that norm implements are acceptable. She knows the norm as a result of growing up in (or become acculturated to) a particular society. As the sociologists Peter Berger and Thomas Luckmann emphasize in their foundational work, *The Social Construction of Reality*:

> In the common stock of knowledge there are standards of role performance that are accessible to all members of a society, or at least to those who are potential performers of the roles in question. This general accessibility is itself part of the same stock of knowledge; not only are the standards of role X generally known, but it is known that these standards are known. Consequently, every putative actor of role X can be held responsible for abiding by the standards, which can be taught as part of the institutional tradition and used to verify the credentials of all performers and, by the same token, serve as controls.[93]

[93] PETER L. BERGER & THOMAS LUCKMANN, THE SOCIAL CONSTRUCTION OF REALITY: A

How does Victoria know that the norm-implemented tradeoff is acceptable? We assume that informational coordination norms are acceptable provided they evolve in sufficiently competitive market conditions under appropriate normative and regulatory constraints. As long as Victoria has reason to think such conditions hold, she knows that the norms are acceptable. We emphasize that norms may be acceptable in this sense but inconsistent with an individual's or group's values. Acceptability (as we use the term) is a minimal standard. Tension between acceptable norms and other values is an important dynamic that motivates critique and social change. The point to emphasize is that Victoria fulfills the *Knowledge Condition* without needing to expend any effort to determine what information the bookstore collects, nor what it does with it. She knows that *whatever* it does, it is acceptable.

Can one replicate this result for, for example, a 25,000 variable, constantly updated e-score system?

C. Predictive Analytics, Norms, and the Knowledge Condition

That would require informational coordination norms that govern the use of predictive analytics, and such shared norms generally do not exist. The existence of such a norm requires that the parties coordinate to create a selective flow information in order to achieve a shared goal, and such goals typically do not exist. Consider insurance companies that set premiums using "your credit history, whether you use department-store or bank credit cards, and even your TV provider."[94] The insurer's goal is to maximum their profits. Do the customers share that goal, or some other relevant goal, with the companies? That is unlikely. Few will know how insurance companies use predictive analytics, and even if they did, the intense privacy debates over the use of predictive analytics would be sufficient to show that there is no agreement on a relevant shared goal.

How should public policy proceed given that uses of predictive analytics are often not governed by relevant informational coordination norms? One attractive answer is to create norms that ensure that the *Knowledge Condition* is fulfilled while also regulating the features in the profile in ways that ensure that human predictive systems do not exhibit the objectionable features of AI.

TREATISE IN THE SOCIOLOGY OF KNOWLEDGE 74 (1967 ed. 1967).

[94] *Special Report: Car Insurance Secrets*, *supra* note 77.

IV. CREATING NORMS

One must do two things to create a coordination norm: first, ensure that people conform to a behavioral regularity; second, ensure that they do so in part because they think they ought to as long as others do in order to realize a shared goal. A combination of education and regulation can achieve those ends.

A. Norm Creation Through Education and Regulation

Anti-littering campaigns are a good example of how education and regulation can create a coordination norm.[95] The coordination norm involved is not an *informational* coordination norm, but we will show how to adapt the process to informational norms. In the early 1950s, almost everyone littered even though almost everyone desired a litter-free environment, but as long as everyone littered, individually taking the time and effort to use waste receptacles would not make the environment cleaner, and people preferred littering to expending pointless time and effort.[96] An intensive advertising campaign combined with legal liability led to a non-littering coordination norm.[97] It convinced people they ought not to litter in order to realize the shared goal of a cleaner environment, and, for that reason, people generally began to use and expect others to use waste receptacles. Littering is one of many examples of the creation of coordination norms. The Nobel Prize winning economist Elinor Ostrom[98] and the philosopher Christina Bicchieri[99] discuss a number of examples.

We propose a similar process of education and regulation for predictive analytics. The proposed norm creation strategy is built around the FTC's standard for an unfair business practice: a predictive

---

[95] *See* Bradford Plumber, *The Origins of Anti-Litter Campaigns*, MOTHER JONES (May 22, 2006), http://www.motherjones.com/politics/2006/05/origins-anti-litter-campaigns/. Two additional examples are campaigns against smoking and eating red meat. *See* John C. Catford, Don Nutbeam & Martin C. Woolaway, *Effectiveness and Cost-Benefits of Smoking Education*, 6 J. PUB. HEALTH 264 (1984); Henry W. Kinnucan, Hui Xiao, Chung-Jen Hsia & John D. Jackson, *Effects of Health Information and Generic Advertising on U.S. Meat Demand*, 79 AM. J. AGRIC. ECON. 13 (1997).

[96] Plumber, *supra* note 95.

[97] *Id.*

[98] *See generally* ELINOR OSTROM, GOVERNING THE COMMONS: THE EVOLUTION OF INSTITUTIONS FOR COLLECTIVE ACTION (reissue ed. 2015) (discussing the importance of norms in fostering cooperation); ELINOR OSTROM, UNDERSTANDING INSTITUTIONAL DIVERSITY (2005) (also discussing the importance of norms in fostering cooperation); Elinor Ostrom, *Collective Action and the Evolution of Social Norms*, 14 J. ECON. PERSP. 137 (2000).

[99] *See generally* CRISTINA BICCHIERI, NORMS IN THE WILD: HOW TO DIAGNOSE, MEASURE, AND CHANGE SOCIAL NORMS (2016) (reporting the results of empirical observation of the evolution of norms).

system should not cause or be likely to cause "substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition."[100] The first step is to explain the relevant notion of a norm.

B. Creation Norms for Predictive Analytics

We sketch an educational and regulatory procedure for creating a norm covering the use of proxies (such as credit scores) in setting auto insurance premiums and then suggest how to generalize to other features of the profile. A reasonable norm in the auto insurance case would be to use proxies only in ways that are sufficiently predictive of driving safety. Note that being "sufficiently predictive" is not just a matter of the predictive reliability of the proxies. It is also a normative question since the use proxies can make what you do in one area of your life have consequences in another.[101]

We suggest the following procedure as a plausible norm-creation process. First, the FTC should hold companies liable for an unfair business practice for using proxies in ways insufficiently predictive of auto safety. FTC enforcement starts the norm-creation process by ensuring that the following behavioral regularity obtains at least some auto insurance companies set premiums using sufficiently predictive proxies, and at least some consumers purchase polices with such premiums. FTC decisions help define what counts as "sufficiently predictive."

To create the norm from that point, one needs to do three more things. First, extend the regularity to (almost) all insurance companies and (almost) all consumers buying auto insurance. Second, ensure that companies and consumers share the goal of companies offering and consumers buying polices based on sufficiently predictive proxies. Third, ensure that, in order to realize that goal, companies and consumers conform to the regularity because they think they ought to conform as long as enough others do.

The first step toward realizing these goals is to convince consumers that insurance companies ought to use only sufficiently predictive proxies. Thus, through educational campaigns, convince consumers that insurance companies should use only sufficiently

---

[100] 15 U.S.C. § 45(n) (2012).

[101] *See supra* section II.D.1.

predictive proxies.[102] As a result, consumers begin to demand insurance premiums based on such proxies.

While it is not inevitable, this combination could lead to insurance companies responding by offering such premiums. They think they ought to in order to meet consumer demand (and thereby ensure adequate revenue).

Eventually, the informational coordination norm exists. First, there is a behavioral regularity; insurance companies use only sufficiently predictive proxies. Second, companies and customers conform to the regularity because think they ought to in order realize the shared goal of using only sufficiently predictive proxies, and third, they trust each other to do so.

One could create norms for the other parts of the profile in similar ways. A plausible norm for the "feedback" part would be that there should be adequate error-correcting feedback. For "data preparation," one norm would be that the information in the database should be sufficiently predictive, where "sufficiently predictive" is again in part a normative notion. For the "predictive target" part of the profile, we suggest a norm requiring predictive targets to meet appropriate standards of fairness, explainability, lack of opacity, and feasibility. For the accuracy part of the profile, we suggest that a norm that requires that there be a sufficiently reliable correlation between a systems predictor data and its output data.

CONCLUSION

We could create norms in the way suggested, but it would require a significant commitment of resources to educational campaigns and FTC regulation. Does society have the political and social will to do that? If not, it does not mean that norms will not evolve. Coordination norms evolve in response to repeated interactions in which the parties give and take from each other.[103] The danger is that, as norms evolve, habituation to current business practices will lead people to accept what now seems objectionable, and norms will arise that give businesses very wide latitude in the use of predictive analytics.[104] Now is the time to intervene in that process.

---

[102] Government, consumer advocate groups, and industry organizations may conduct the advertising campaigns.

[103] *See generally* EDNA ULLMANN-MARGALIT, THE EMERGENCE OF NORMS (Oxford Univ. Press 1st ed. 2015).

[104] On habituation, *see* BERGER & LUCKMANN, *supra* note 93, at 53-58.