McGill Law Journal
Revue de droit de McGill

Strategic Games and Algorithmic Secrecy

Ignacio N. Cofone and Katherine J. Strandburg

Volume 64, Number 4, June 2019

Programming Governance/Governing Programming: Regulatory Challenges on the Edge of Technology

URI: https://id.erudit.org/iderudit/1074151ar
DOI: https://doi.org/10.7202/1074151ar

See table of contents

Publisher(s)

McGill Law Journal / Revue de droit de McGill

ISSN

0024-9041 (print)
1920-6356 (digital)

Explore this journal

Cite this article

Cofone, I. N. & Strandburg, K. J. (2019). Strategic Games and Algorithmic Secrecy. *McGill Law Journal / Revue de droit de McGill*, *64*(4), 623–663. https://doi.org/10.7202/1074151ar

Article abstract

We challenge a claim commonly made by industry and government representatives and echoed by legal scholarship: that algorithmic decision-making processes are better kept opaque or secret because otherwise decision subjects will "game the system", leading to inaccurate or unfair results. We show that the range of situations in which people are able to game decision-making algorithms is narrow, even when there is substantial disclosure. We then analyze how to identify when gaming is possible in light of (i) how tightly the decision-making proxies are tied to the factors that would ideally determine the outcome, (ii) how easily those proxies can be altered by decision subjects, and (iii) whether such strategic alterations ultimately lead to mistaken decisions. Based on this analysis, we argue that blanket claims that disclosure will lead to gaming are over-blown and that it will often be possible to construct socially beneficial disclosure regimes.

McGill Law Journal — Revue de droit de McGill

STRATEGIC GAMES AND ALGORITHMIC SECRECY

*Ignacio N. Cofone and Katherine J. Strandburg**

We challenge a claim commonly made by industry and government representatives and echoed by legal scholarship: that algorithmic decision-making processes are better kept opaque or secret because otherwise decision subjects will "game the system", leading to inaccurate or unfair results. We show that the range of situations in which people are able to game decision-making algorithms is narrow, even when there is substantial disclosure. We then analyze how to identify when gaming is possible in light of (i) how tightly the decision-making proxies are tied to the factors that would ideally determine the outcome, (ii) how easily those proxies can be altered by decision subjects, and (iii) whether such strategic alterations ultimately lead to mistaken decisions. Based on this analysis, we argue that blanket claims that disclosure will lead to gaming are over-blown and that it will often be possible to construct socially beneficial disclosure regimes.

Nous contestons une affirmation couramment véhiculée par certains représentants de l'industrie et des gouvernements et qui est parfois reprise dans la littérature juridique, soit l'idée selon laquelle il est préférable de garder les processus de décision algorithmiques opaques ou secrets, faute de quoi les sujets des décisions pourraient « déjouer le système », ce qui conduirait à des résultats inexacts ou injustes. Nous montrons que l'éventail des situations dans lesquelles les gens sont capables de se servir des algorithmes de prise de décision pour son propre avantage est étroit, même lorsqu'une quantité substantielle d'information a été divulguée. Nous analysons ensuite comment discerner les situations où il est possible de détourner les algorithmes en fonction (i) du degré de connexion entre les indicateurs de décision et les facteurs qui, idéalement, détermineraient le résultat, (ii) de la facilité avec laquelle ces indicateurs peuvent être modifiés par les sujets des décisions et (iii) de la possibilité que ces modifications stratégiques conduisent finalement à des décisions erronées. Sur la base de cette analyse, nous soutenons que les allégations générales selon lesquelles la divulgation conduirait à des détournements sont exagérées et qu'il sera souvent possible de mettre en place des régimes de divulgation socialement bénéfiques.

* Ignacio N. Cofone, Assistant Professor, McGill University Faculty of Law. Katherine J. Strandburg, Alfred Engelberg Professor of Law, NYU School of Law. We thank Sebastian Benthall, Cassi Carley, Alessa Dassios, Guiseppe Dari-Mattacci, Ashley Gorham, Yafit Lev-Aretz, John Nay, Garry Gabison, Mason Marks, Julia Powles, Ira Rubenstein, Madelyn Sanfilippo, Mark Verstraete and Ari Waldman for their helpful comments. The paper also benefited from numerous comments on presentations at the NYU Privacy Research Group, NYU Information Law Institute, and the Programming Governance & Governing Programming Conference at McGill University. Katherine Strandburg acknowledges the generous support of the Filomen D'Agostino and Max E. Greenberg Research Fund. Ignacio Cofone is grateful for the support of the Social Sciences and Humanities Research Council (Insight Development Grant) and the Canadian Institute for the Administration of Justice Charles D. Gonthier Research Fellowship. We also thank Malaya Powers for her outstanding research assistance.

Citation: (2019) 64:4 McGill LJ 623 — Référence : (2019) 64:4 RD McGill 623

Introduction

Algorithmic disclosure involves an uneasy trade-off between accountability and effectiveness. Many scholars have pointed out the value of disclosing algorithmic decision-making processes to promote accountability and procedural fairness, and to ensure people's civil rights.[1] Other scholars and policy-makers respond that disclosure would undermine decision-making effectiveness and fairness by enabling decision subjects to "game the system" to achieve undeserved beneficial outcomes.[2] This article analyzes when such concerns about "gaming" are plausible. More specifically, we aim to determine under what conditions revealing information to subjects about how an algorithmic decision is made may lead to inefficient or unfair results.

"Gaming the system" is not a new concern. There is a robust literature, particularly in game theory, analyzing when decision subjects can game decision makers. The growing use of big data and machine learning has drawn renewed attention to the "gaming" issue and introduces important nuances.[3] In particular, while some types of machine-learning-based algorithms[4] have a degree of inherent opacity,[5] the threat of "gam-

---

1 See e.g. Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System" (2018) 70:5 Stan L Rev 1343; Sonia K Katyal, "Private Accountability in the Age of Artificial Intelligence" (2019) 66:1 UCLA L Rev 54 at 120. See also Cary Coglianese & David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017) 105:5 Geo LJ 1147 at 1184 for a nuanced discussion about the specific circumstances under which automated machine learning will constitute a violation of an individual's constitutional right to due process.

2 See e.g. Joshua A Kroll et al, "Accountable Algorithms" (2017) 165:3 U Pa L Rev 633 at 639; Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York: New York University Press, 2017) at 136; Jane Bambauer & Tal Zarsky, "The Algorithm Game" (2018) 94:1 Notre Dame L Rev 1 at 15.

3 In the literature, gaming presupposes that as soon as a model is disclosed, interested parties may try to detect the proxies involved and alter them, thus considerably diminishing the value of the model. Gaming therefore commonly refers to behaviour that intentionally alters the proxies at issue. See Paul B de Laat "Big Data and Algorithmic Decision Making" (2017) 47:3 ACM Computers & Society 39 at 48.

4 See Ian Goodfellow, Yoshua Bengio & Aaron Courville, *Deep Learning* (Cambridge, Mass: MIT Press, 2016) at 8.

5 See Danielle Keats Citron, "Technological Due Process" (2007) 85:6 Wash UL Rev 1249 at 1280; Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms" (2016) 3:1 Big Data & Society 1 at 4. It should be noted, however, that even in opaque deep learning algorithms there is always something to disclose. Therefore, the algorithmic secrecy concern, and the reasons to bracket it given in this article, are applicable to purposeful obfuscation across different types of technology.

ing" is often used by both governments and private entities to justify purposeful secrecy about aspects that could feasibly be disclosed.[6]

Fundamentally, the gaming threat stems from decision maker reliance on proxies for decision criteria. But proxies, though imperfect, are a necessary, but not sufficient, condition for socially undesirable gaming. Socially undesirable gaming occurs only when strategic behaviour by decision subjects makes the proxy less informative in relation to the decision maker's underlying goals. Socially undesirable gaming means that decision makers will make less optimal decisions or must engage in further screening efforts.

Many proxies are not gameable, as a practical matter, or are gameable only in some respects. Having established that the use of relatively imperfect proxies is a crucial prerequisite for gaming, we describe three tiers of proxies involved in machine-learning-based algorithms. We then identify the limited levers available for decision subjects seeking to exploit proxies that are loosely tied to the ideal decision-making criteria to affect decision outcomes. In most contexts, the only handle decision subjects can use to affect such an algorithm's output is the data about decision subject characteristics or "features" it uses as input. Since decision subjects are rarely able to hack into that data to falsify it directly, the only strategy available to them is to alter the behaviour that is reflected in that data.

We thus employ concepts based on signaling theory[7] to create a framework for assessing the extent to which it is practically feasible for decision subjects to alter their features to obtain a more beneficial outcome from a decision-making algorithm. We then argue that, even where decision subjects can feasibly alter their behaviour to obtain more favourable outcomes, the result may not amount to socially undesirable gaming. Decision subjects' strategic behaviour can be socially beneficial if it makes them more truly deserving of a beneficial decision, or if it corrects errors

---

6 See Rebecca Wexler, "When a Computer Program Keeps You in Jail", *The New York Times* (20 January 2018), online: <www.nytimes.com> [perma.cc/354L-MSMT]. Algorithmic secrecy is often bolstered or effectuated by asserting trade secrecy. There is an important ongoing scholarly debate about whether trade secrecy is a valid or effective mechanism for promoting innovation in decision-making algorithms. The plausibility of gaming clearly affects the balance of social costs and benefits associated with assertions of trade secrecy, but we leave that question aside for now.

7 Signaling theory is a branch of game theory that studies how agents strategically convey a signal to a principal in situations of asymmetric information. See e.g. Michael Spence, "Job Market Signaling" (1973) 87:3 QJ Economics 355 at 356–57; Douglas G Baird, Robert Gertner & Randal Picker, *Game Theory and the Law*, revised ed (Cambridge, Mass: Harvard University Press, 1998) at ch 4; Nolan McCarty & Adam Meirowitz, *Political Game Theory: An Introduction* (Cambridge, UK: Cambridge University Press, 2014) at ch 8.

caused by an algorithm's predictive inaccuracy. A framework for assessing the circumstances in which disclosure will and will not lead to socially undesirable gaming may help policy-makers and scholars take a nuanced approach to algorithmic disclosure.

We emphasize throughout that delving into the specifics of gameability for particular algorithms is important for disclosure policy because disclosure is not an all-or-nothing matter. Even if complete disclosure raises plausible concerns about gaming, decision makers can usually disclose significant aspects of their algorithms without triggering socially undesirable gaming. Such limited disclosures may often provide substantial tools for algorithm accountability. Finally, we discuss the power that decision makers often have to create more or less gameable algorithms and the strategic considerations that might drive them to game the gameability threat.

Our analysis suggests that, from a social perspective, the threat from "gaming" is overstated by its invocation as a blanket argument against disclosure. The consequential over-secrecy deprives society not only of the benefits of disclosure to decision subjects, but also of the improvements in decision quality that could result when disclosure improves accountability. Policy debates should thus focus less on *whether* to require disclosure and more on *what information* should be disclosed.

Part I of this article illustrates the common use of the threat of "gaming the system" as a rationale for non-disclosure. Part II describes the relationships between decision-making proxies and gaming and identifies the limited opportunities for gaming offered by algorithms created using big data and machine learning. It concludes that, in most contexts, decision subjects' only avenue for gaming is to strategically alter the features that the algorithm uses as inputs. Part III analyzes, as a practical matter, how the properties of the input feature set used by a decision-making algorithm affect the algorithm's gameability. It highlights how the complexity commonly associated with machine-learning decision algorithms is likely to hinder gaming, and then distinguishes socially undesirable gaming from strategic behaviour that benefits both decision subjects and society as a whole. Finally, Part IV briefly considers the role of decision makers in the gaming/disclosure trade-off. It points out that decision makers often have considerable control not only over what is disclosed about their decision algorithms, but also over the gameability of those algorithms. Decision makers can thus respond to the potential for decision subject gaming either through secrecy or by investing in less gameable algorithms. Decision makers are also strategic actors, however, whose choices are driven by private interests that align imperfectly with social interests. The article concludes by summarizing the prerequisites for a plausible gaming threat and discusses how disclosure regimes might be constructed to minimize such threats.

I. The Ubiquitous Gaming Trope

This issue arises from an especially common scenario between decision subjects and decision makers: decision subjects (or potential decision subjects) often demand information about the bases for decisions that disadvantage them. Decision makers respond that disclosure is "undesirable, such as when it discloses private information or permits tax cheats or terrorists to game the systems determining audits or security screening."[8] The force of the gaming argument depends, by implication, on an assertion that the social costs of gaming outweigh the benefits of disclosure. The spectre of gaming is raised in a range of situations, yet the implied cost-benefit analysis is rarely spelled out in any detail.[9] This Part provides a brief overview of some situations in which the threat of "gaming the system" has been offered to forestall disclosure of decision-making criteria. It then describes some wrinkles in the way that the gaming argument plays out for automated decision-making algorithms.

A. Governmental Opacity

Over 10 years ago, William Stuntz said in a much-cited article that

> [l]aw enforcement is a game of cat and mouse. The government makes its move, criminals respond, government adapts, and the game goes on. Thankfully, most criminals are not too bright, so the

---

8 See Joshua A Kroll et al, "Accountable Algorithms" (2017) 165 U Pa L Rev 633 at 633–34, 638. See also *ibid* at 639 where the authors state that

> [t]he process for deciding which tax returns to audit, or whom to pull aside for secondary security screening at the airport, may need to be partly opaque to prevent tax cheats or terrorists from gaming the system. When the decision being regulated is a commercial one, such as an offer of credit, transparency may be undesirable because it defeats the legitimate protection of consumer data, commercial proprietary information, or trade secrets. Finally, when an explanation of how a rule operates requires disclosing the data under analysis and those data are private or sensitive (e.g., in adjudicating a commercial offer of credit, a lender reviews detailed financial information about the applicant), disclosure of the data may be undesirable or even legally barred.

9 One contribution in this area has analyzed the circumstances in which the benefits of an explanation outweigh the costs from a moral, social, and legal point of view. See Finale Doshi-Velez & Mason Kortz, "Accountability of AI Under the Law: The Role of Explanation" (2017) Berkman Klein Center for Internet & Society Working Paper at 4–5, online: *Digital Access to Scholarship at Harvard* <dash.harvard.edu> [perma.cc/C822-SR8E] (the authors suggest that the situations under which an explanation is necessary are as follows: the decision must have been acted on in a way that has an impact on a person other than the decision-maker, there must be value to knowing if the decision was made erroneously, and there must be some reason to believe that an error has occurred (or will occur) in the decision-making process, which could be a result of distrust in the integrity of the system).

> game is easily won. Terrorists are different; the most dangerous ones are smart and well-motivated. Whatever information they have, they will use. There is something deeply crazy about publicly debating what law-enforcement tactics the government should use to catch people who are happily listening to the debate and planning their next moves.[10]

Arguments against algorithmic disclosure in the law enforcement context are the latest incarnation of this view. A recent article argues that both the process and criteria for algorithms used for purposes like tax auditing and terrorism prevention must be opaque in order to avoid gaming and introducing new risks. It contends:

> A major problem is that the public interest disclosure of just algorithms might be likely to produce serious negative consequences. On many platforms the algorithm designers constantly operate a game of cat-and-mouse with those who would abuse or "game" their algorithm. These adversaries may themselves be criminals (such as spammers or hackers) and aiding them could conceivably be a greater harm than detecting unfair discrimination in the platform itself.[11]

Though more skeptical that law enforcement can win the game, Andrew Ferguson similarly observes that

> people committed to criminal activity will learn how to outsmart additional police surveillance. If automobiles are monitored through ALPR, criminal actors will stop using their own cars. If cellphones are intercepted, criminal actors will use disposable phones. The cat-

---

[10] William J Stuntz, "Secret Service: Against Privacy and Transparency", *The New Republic* (17 April 2006) at 14–15 (adding that "[i]n order to govern wisely, the government should know as much as possible about those it governs. And the citizenry should know a lot less about government officials" at 12). See also Douglas Martin, "W.J. Stuntz, Who Stimulated Legal Minds, Dies at 52", *The New York Times* (20 March 2011), online: <www.nytimes.com> [perma.cc/66W3-MJGT] ("Mr. Stuntz wrote for newspapers and magazines on issues beyond the law. In an article in The New Republic in 2006, he raised liberal eyebrows by saying that government could be more effective in fighting terrorism if it were less transparent and more concerned with protecting its own privacy than that of its citizens").

[11] Christian Sandvig et al, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms" (Paper delivered at the Annual Meeting of the International Communication Association, Seattle, 22 May 2014) [unpublished] at 9 (also stating that "[f]or example, a select few Internet platforms are already open about their algorithms, typically because they subscribe to a culture of openness influenced by the open source movement in software engineering. One such platform is Reddit (whose code is open source). However, to prevent spambots from using the disclosed algorithm to attack Reddit, making its rating and commenting systems useless, a kernel of the algorithm (called the 'vote fuzzing' code) must remain closed source and secret, despite Reddit's aspirations to transparency" at 9).

> and-mouse game of cops and robbers will continue no matter the technology.[12]

Law enforcement authorities often justify hiding their use of Sting-Rays and Wolfhounds—controversial cellphone tracking devices—by arguing that disclosing their use of the technology or releasing information about how it works would undermine its value by helping targets to evade it.[13] The chief of the FBI tracking technology unit, Bradley Morrison, has stated in an affidavit that "[t]he FBI routinely asserts the law enforcement sensitive privilege over cell site simulator equipment because discussion of the capabilities and use of the equipment in court would allow criminal defendants, criminal enterprises, or foreign powers, should they gain access to the items, to determine the FBI's techniques, procedures, limitations, and capabilities in this area"[14] and that "[t]his knowledge could easily lead to the development and employment of countermeasures to FBI tools and investigative techniques by subjects of investigations and completely disarm law enforcement's ability to obtain technology-based surveillance data in criminal investigations."[15] Spokesperson for the Baltimore County Police, Elise Armacost, similarly argues that "[w]e can't disclose any legal requirements associated with the use of this equipment ... [because] [d]oing so may disclose how we use it, which, in turn, interferes with its public-safety purpose."[16]

The argument that disclosure will lead to gaming is also used in resisting discovery in litigation challenging police practices: "When plaintiffs asked a New York court to turn over NYPD stop-and-frisk data, the Department objected on the grounds that this would 'give away information about specific policing methods, such as location, frequency of stops, and patterns.'"[17] Addressing law enforcement's use of big data, legal scholar Tal Zarsky argues that "[t]hose striving to game the law enforcement process will greatly benefit from insights into the aggregation and collation process. They will use such information to understand how they

---

12 *Supra* note 2 at 184. Ferguson also writes, "[t]he second reason why demanding algorithmic transparency may be misguided is that as a technological matter, it may be impossible. ... Revealing the source code means revealing the company's competitive advantage in business" (*ibid* at 138).

13 See Barry Friedman, "Secret Policing" [2016] U Chicago Leg Forum 99 at 103, 108–09, 120–21.

14 Cyrus Farivar, "FBI Would Rather Prosecutors Drop Cases Than Disclose Stingray Details", *Ars Technica* (7 April 2015), online: <arstechnica.com> [perma.cc/NGV6-VSPP].

15 *Ibid*.

16 Jennifer Valentino-DeVries, "Police Snap Up Cheap Cell Phone Trackers", *The Wall Street Journal* (19 August 2015), online: <www.wsj.com> [perma.cc/R97S-WNEC].

17 Friedman, *supra* note 13 at 119–20.

might be able to escape having their information aggregated into one dataset."[18]

The claim that law enforcement strategies must be shielded from public inquiry to avoid circumvention is also enshrined in exemption 7(E) of the *Freedom of Information Act*, which permits the refusal of requests that "would disclose techniques and procedures for law enforcement investigations or prosecutions, or would disclose guidelines for law enforcement investigations or prosecutions if such disclosure could reasonably be expected to risk circumvention of the law."[19]

Tax auditing also relies on predictive algorithms: the Internal Revenue Service (IRS) evaluates which tax returns to audit using a scoring algorithm called the Purpose of Discriminant Inventory Function (DIF). The IRS shrouds its prediction models for targeting its auditing efforts in complete secrecy.[20] Indeed, while individuals can learn their credit scores, the IRS keeps not only its algorithm, but also the DIF scores secret. The argument for maintaining such secrecy is that any degree of disclosure will facilitate tax evasion going undetected.[21] "The simplest way to understand this argument is that knowledge of the inner workings of the automated prediction models in the hands of adversaries will allow them to 'game the system.'"[22] This is, the argument goes, because machines, unlike humans, will continue using a certain procedure in a way that makes them more predictable.[23]

As Zarsky points out, referencing Harcourt:

> if the IRS focuses auditing on individuals who meet specific criteria and such information becomes public, individuals who are not part of this group will alter their conduct and cheat on their taxes in greater numbers, as they understand they can do so without being

---

18 See Tal Z Zarsky, "Transparent Predictions" (2013) 2013:4 U Ill L Rev 1503 at 1564.

19 5 USC § 552(b)(7)(E) (2012). See also "United States Department of Justice Guide to the Freedom of Information Act" (last modified 7 March 2019) at Exemption 7(E), online: *United States Department of Justice* <www.justice.gov> [perma.cc/E69T-7X9K]. Note that a similar understanding of Exemption 2 was rejected in *Milner v Department of the Navy*, 562 US 562 (2011).

20 See Zarsky, *supra* note 18 at 1510–12.

21 See *ibid* at 1512, 1553–54.

22 *Ibid* at 1554.

23 See *ibid* at 1554; Ignacio N Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21:2 Stan Tech L Rev 167 at 183–86 (discussing how algorithms have different levels of unpredictability depending on the technology). See also Subpart II.D., *below*.

> detected. Therefore, (transparent) predictive modeling leads to more, not less, crime.[24]

Judges have also accepted this line of argument, suggesting that releasing people's DIF scores could help them circumvent tax law by lowering their scores to avoid audit.[25] The IRS also uses automated algorithms to detect evasion; these algorithms are also kept secret to forestall gaming.[26]

B. Private Opacity

The "gaming the system" argument is not exclusive to government actors; private companies utilize it with equal frequency.[27] For example, the argument is commonly advanced in discussions about credit scores.[28] Credit card companies develop credit risk scoring based on various behavioural metrics.[29] These companies protect their algorithms with trade secrets and invoke two sorts of arguments against revealing them: that it would diminish innovation and that it would enable people to game the predictive algorithm, reducing its accuracy.[30] "Credit bureaus will object

---

[24] Bernard E Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age* (Chicago: University of Chicago Press, 2007) at 23. See also Zarsky, *supra* note 18 (paraphrasing Harcourt to say that "if the IRS focuses auditing on individuals who meet specific criteria and such information becomes public, individuals who are not part of this group will alter their conduct and cheat on their taxes in greater numbers, as they understand they can do so without being detected. Therefore, (transparent) predictive modeling leads to more, not less, crime" at 1558).

[25] See e.g. *Huene v US Department of the Treasury et al*, (ED Cal Dist Ct 2012) No 2:11-cv-02109 JAM KJN PS at 7 (recommendation for granting summary judgment).

[26] See Harcourt, *supra* note 24 at 9; Zarsky, *supra* note 18 at 1554–58.

[27] Given the nature of algorithmic secrecy for private sector uses, some have suggested that it should not be up to the state to address issues of algorithmic accountability. Rather, we are better to look to the private industry to play the dominant role in algorithmic accountability and transparency. See Katyal, *supra* note 1 at 61.

[28] For a detailed example explaining how an automated credit scoring system functions using machine learning, see Sandra Wachter, Brent Mittelstadt & Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2017) 7:2 Intl Data Privacy L 76 at 78.

[29] See Danielle Keats Citron & Frank Pasquale, "The Scored Society: Due Process for Automated Predictions" (2014) 89:1 Wash L Rev 1 at 4–5; Deval L Patrick, Robert M Taylor III & Sam SF Caligiuri, "The Role of Credit Scoring in Fair Lending Law: Panacea or Placebo?" (1999) 18 Annual Rev Banking L 369 at 370.

[30] See Citron & Pasquale, *supra* note 29 ("[t]here is also not adequate evidence to give credence to 'gaming' concerns—i.e., the fear that once the system is public, individuals will find ways to game it. While gaming is a real concern in online contexts, where, for example, a search engine optimizer could concoct link farms to game Google or other ranking algorithms if the signals became public, the signals used in credit evaluation are far costlier to fabricate" at 26); Brenda Reddix-Smalls, "Credit Scoring and Trade Secrecy: An Algorithmic Quagmire or How the Lack of Transparency in Complex Fi-

that transparency requirements—of any stripe—would undermine the whole reason for credit scores. Individuals could 'game the system' if information about scoring algorithms were made public or leaked in violation of protective orders. Scored consumers would have ammunition to cheat, hiding risky behavior."[31] These arguments resist public calls, including from the Occupy Wall Street movement, for greater algorithmic transparency.[32]

The narrative of gaming and counter-gaming is also common in fraud detection. For credit card fraud detection, for example, it has been said that "it is safe to assume that the thief is likely to attempt a fraudulent transaction with a belief that his transaction may be monitored on the basis of transaction amount"[33] and "the loss can be minimized if the system is able to predict the next move of the thief correctly. ... The dilemma for the thief, on the other side, is to be able to choose a transaction range that has not been predicted by the [Fraud Detection System]."[34] Companies' credit card fraud detection methods are therefore hidden from the

---

nancial Models Scuttled the Finance Market" (2011) 12:1 UC Davis Bus LJ 87 ("[t]hese algorithmic credit scoring models upset the balance of asymmetric information for the consumer. The lack of information does not allow the consumer to play the game fairly" at 119). See also Christer Holloman, "Your Facebook Updates Now Determine Your Credit Score", *The Guardian* (28 August 2014), online: <www.theguardian.com> [perma.cc/PUU2-J9UE] ("[f]or obvious reasons – and not least because they want to avoid having consumers trying to game the system – none of the companies I spoke to could reveal the exact details of their process").

31 Citron & Pasquale, *supra* note 29 at 30.

32 See Odysseas Papadimitriou, "Occupy Wall Street Is Only Half Right About Credit Reform", *Time* (9 April 2012), online: <business.time.com> [perma.cc/7FQL-3XS6]. See also Kaveh Waddell, "How Algorithms Can Bring Down Minorities' Credit Scores", *The Atlantic* (2 December 2016), online: <www.theatlantic.com> [perma.cc/6SKC-AYMB] (there is also a push for greater transparency about credit score algorithms on the basis that when algorithms analyze people's social connections, the algorithms may be engaging in credit discrimination. Therefore, "more types of information can help people who lack credit scores, or who might not have the usual indicators of creditworthiness, access loans and bank accounts that might otherwise be closed off to them").

33 Vishal Vatsa, Shamik Sural & AK Majumdar, "A Game-Theoretic Approach to Credit Card Fraud Detection" (Paper delivered at the International Conference on Information Systems Security, Kolkata, 19–21 December 2005) in Sushil Jajodia & Chandan Mazumdar, eds, *Lecture Notes in Computer Science*, vol 3803 (Berlin: Springer-Verlag, 2005) at 267.

34 *Ibid* (adding that "the game being played between the FDS and the fraudster is one of incomplete information since the fraudster would be completely unaware of the modus operandi of the Detection System. ... since we assume that the situation is one of repeated games, the fraudster can use his past experience to build upon his belief about the FDS strategy" at 269.)

public to avoid alleged gaming.[35] Those methods also are often modified to prevent "criminals [from] adapting to their strategies."[36]

A less well-known context for assertions that secrecy prevents gaming is insurance. The purchase of insurance is essentially a game of chance, in which policy holders take an upfront financial loss to avoid a (potentially larger) future loss.[37] Historically, the analogy to gambling was even more apt: "life insurance served as a vehicle for gaming. Though ostensibly devoted to risk avoidance, life insurance arose from and drew much of its initial popularity in eighteenth-century England from people's taste for gambling on others' lives."[38] Insurance pricing sets the terms of the gamble. Insurers that engage in dynamic pricing hide the methods through which they arrive at different premiums for different policyholders, and many also hide the scores that determine those premiums.[39] When the Consumers Union attempted to get access to scoring models used by a group of insurers (AIG, Liberty Mutual, Nationwide, and State Farm), the insurers prevailed by claiming that the models were trade secrets and that keeping them confidential was in the public interest.[40]

II. What Can Be Gamed

A decision-making rubric (whether automated or not) is gameable only if a decision subject can take actions that improve the likelihood of a beneficial decision without changing her underlying qualifications for a beneficial outcome. As Jane Bambauer and Tal Zarsky put it in the context of decision-making algorithms, "gaming involves a change in the subject's behaviour to affect the algorithm's estimate *without causing any change to the key characteristic* that the algorithm is attempting to measure."[41] As we discuss in this Part, gaming is not an inevitable response to disclosure but can occur only when certain key prerequisites are met.

---

35 See Yufeng Kou et al, "Survey of fraud detection techniques" (IEEE delivered at the International Conference on Networking, Sensing & Control, Taipei, 21–23 March 2004) 749 at 753.

36 See *ibid* at 749.

37 See Pat O'Malley, "Imagining Insurance: Risk, Thrift, and Life Insurance in Britain" in Tom Baker & Jonathan Simon, eds, *Embracing Risk: The Changing Culture of Insurance and Responsibility* (Chicago: University of Chicago Press, 2002) 97 at 111.

38 Geoffrey Clark, "Embracing Fatality Through Life Insurance in Eighteenth-Century England" in Baker & Simon, *ibid*, 80 at 80.

39 See "Caution! The Secret Score Behind Your Auto Insurance" (2006) at 46, online (pdf): *Consumer Reports* <advocacy.consumerreports.org> [perma.cc/8KZK-R2F5].

40 See *ibid*.

41 *Supra* note 2 at 10.

Specifically, imperfect proxies,[42] which are loosely connected to the ideal decision criteria and can be affected by decision subjects, are prerequisites for gaming. Moreover, socially undesirable gaming can only occur when decision subjects are able to exploit those loose connections to obtain beneficial decisions that they do not deserve. This Part explores what parts of a decision-making algorithm can be gamed.

A. Gaming Depends on Proxies

Decision-making proxies are used when the ideal decision-making criteria are unascertainable as a practical matter or simply unknowable.[43] Both these situations are commonplace. The ideal criteria may be unascertainable because subjects have no way to credibly communicate them or because they choose to obfuscate them. Many decisions, in contexts varying from employment to school admission to pre-trial detention, would ideally be based on the decision subject's future behaviour, an inherently unknowable quality. For example, employers must employ proxies to attempt to predict which candidates are most likely to perform well if they are hired. A candidate's previous performance in a very similar job is likely to be closely tied to future job performance and is thus an accurate proxy, but it is only a proxy. Moreover, it can be employed only when credible information about past performance is available; this information may be difficult to obtain and is unavailable for candidates without prior experience. To cope with these limitations, employers regularly use proxies based on available information, such as educational background, length and types of prior experience (rather than previous performance), test scores, an interviewer's sense of the candidate's energy level, intelligence, and other intangible qualities.[44]

The choice of proxies thus defines the size and nature of the gap between the outcome of a decision-making process and the ideal decision.

---

[42] See Ajay Agrawal, Joshua Gans & Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence* (Boston: Harvard Business Review Press, 2018) at 64. See also Bambauer & Zarsky, *supra* note 2 at 7.

[43] See generally Pamela Hogle, "Proxies in eLearning Data Reveal Promise, Pitfalls of AI" (27 August 2017), online: *Learning Solutions* <www.learningsolutionsmag.com> [perma.cc/P5KE-KTTA].

[44] Automated decision-making tools similarly use less accurate proxies to sort through potential job applicants. For example, one program looks at "an applicant's life online, ranking candidates on the creativity, leadership and temperament evidence on social networks and search results" as a proxy for hiring successful job applicants. See Frank Pasquale, *The Black Box: The Secret Algorithms That Control Money and Information* (Cambridge, Mass: Harvard University Press, 2015) at 34. See also Cathy O'Neil, "How Algorithms Rule Our Working Lives", *The Guardian* (1 September 2016), online: <www.theguardian.com> [perma.cc/L9M8-L9HM].

Proxies may be tied to the outcome of ultimate interest to the decision-maker to varying degrees and in different ways. As a result, they vary in their reliability as indicators of the characteristics that matter to a decision maker.

When a decision maker employs an imperfect proxy, a decision subject may be able to fool the decision maker into issuing a beneficial decision by modifying the proxy without changing the underlying characteristics that would lead to a detrimental outcome. Though algorithms will vary in the extent to which they rely on features that can be easily modified by decision subjects to produce more beneficial outcomes, the general rule is that loose proxies are more susceptible to gaming. Generally, the more tightly a decision-making proxy is tied to the ideal decision-making criteria, the more difficult it will be for decision subjects to succeed in fooling the proxy. In other words, strongly tied proxies are more difficult and sometimes effectively impossible to game.[45]

In sum, proxies are ubiquitous in decision making, statistically noisy to varying degrees, and often affected by various sorts of bias. Gaming is possible to the extent that a decision-making proxy is imperfect. The use of loosely tied proxies is thus the first prerequisite for gaming.

B. Three Layers of Proxies

Assessing the threat of gaming requires a basic understanding of how proxies are embedded in decision-making processes. Machine learning is touted primarily as a mechanism for improving on human ability to predict an outcome variable by taking account of patterns in large datasets involving a large number of input features and eliminating calculation errors.[46] The performance of such an algorithm is affected by three layers of proxies,[47] which are probabilistically tied to the characteristics that would ideally be grounds for a decision.[48] These distinct tiers of proxy relation-

---

[45] One exception to this general observation is that it is possible to imagine a proxy that is tightly, but only temporarily, correlated with a characteristic of underlying interest to the decision-maker. For example, a gang hangout may be highly correlated with involvement in drug trafficking until the gang learns that it is under surveillance by the police, at which point the gang might break the correlation by choosing a different meeting place.

[46] See e.g. Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (New York: Basic Books, 2015) at 6–10.

[47] See Agrawal, Gans & Goldfarb, *supra* note 42. See also Bambauer & Zarsky, *supra* note 2 at 7.

[48] In machine learning, proxies are defined as models which substitute for facts. See Hogle, *supra* note 43, citing Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increase Inequality and Threatens Democracy* (New York: Broadway Books, 2017) ("A

ships[49] are illustrated in Figure 1 using a hypothetical algorithm for predicting the risk of recidivism as a factor in parole decisions.[50]

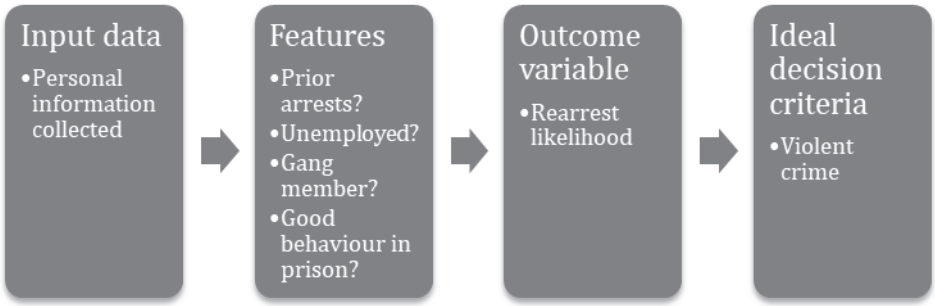

Figure 1: Illustrates proxy relationships between data, features, outcome variables, and decision criteria.

At the first level, the input data is only a proxy for the decision subject's features because the data may be erroneous, biased or incomplete, or the feature labels may not accurately describe the source and meaning of the data.[51] In our hypothetical recidivism algorithm, for example, a decision subject's previous number of arrests may have been mistyped, some of her employment may have been informal and not recorded in a government database used as a source for employment data, or the feature labeled "good behaviour in prison" could mean anything from number of rule infractions to a guard's assessment, both of which could be biased representations of whether the defendant exhibited "good behaviour." The

---

proxy might consist of data about actual people who are similar to the person whose learning needs you're trying to anticipate. Or, it could be information that is easily available or legal to gather, like a person's ZIP code, that proxies information–intentionally or not–that cannot easily or ethically be considered, like race. Most insidiously, proxies can include patterns that an AI (artificial intelligence) algorithm has detected and is using, without the knowledge or intention of the humans using the AI-powered application. ... Proxies are ubiquitous because people building data-based, automated tools 'routinely lack data for the behaviours they're most interested in,' Cathy O'Neil wrote in *Weapons of Math Destruction*.' So they substitute stand-in data, or proxies. They draw statistical correlations between a person's ZIP code or language patterns and her potential to pay back a loan or handle a job'" at 17–18).

49 See generally David Lehr & Paul Ohm, "Playing with the Data: What Legal Scholars Should Learn About Machine Learning" (2017) 51:2 UC Davis L Rev 653.

50 See generally Jessica M Eaglin, "Constructing Recidivism Risk" (2017) 67:1 Emory LJ 59.

51 See generally Lehr & Ohm, *supra* note 49; John Nay & Katherine J Strandburg, "Generalizability: Machine Learning and Humans in the Loop", in Ronald Vogl, ed, *Research Handbook on Big Data Law* (forthcoming 2020), online: <ssrn.com/abstract=3417436> [perma.cc/P8UH-7VKD]. See also Subpart II.A., *below*.

strength of the proxy relationship between input data and features is controlled by the sources the decision-maker relies on for that data.

At a second level, the algorithm uses a decision subject's feature data to compute a predicted value of the outcome variable that is only a proxy for that individual's "true" outcome value. To compute that predicted value, the machine-learning-based algorithm combines the feature data according to a formula that it "learned" by fitting training data about other individuals. The algorithm's output is thus only a probabilistic estimate of the decision subject's true outcome variable value.[52] The strength of this proxy relationship is controlled by the algorithm designer's choices of feature set, outcome variables, training data, machine learning approach and so forth. The hypothetical algorithm in Figure 1, for example, uses such a formula to combine information about a prisoner's prior arrests, prior employment, gang affiliation and behaviour in prison to estimate the outcome variable "likelihood of rearrest if released on parole." The algorithm's output may under- or over-estimate the likelihood that a particular prisoner would be rearrested if released. An algorithm that learned from different source data or used a different set of features might well give a better (or worse) estimate.

Finally, the algorithm's outcome variable is chosen by the algorithm designer as a proxy for some ideal decision criterion. In our hypothetical, the decision maker wishes to base parole decisions on the risk that the defendant will commit a violent crime during the period when he or she would otherwise have remained in prison. The hypothetical algorithm uses the outcome variable "likelihood of rearrest" as a proxy for the risk that the defendant will commit a violent crime.

All three of these levels of proxies degrade the validity of a machine-learning-based decision-making algorithm. Inaccurate input data will lead to erroneous predictions for the outcome variable. Even if the input data is accurate, the algorithm's estimates of the outcome variable can be erroneous because the feature set used in the model is insufficient to account for all characteristics that affect the outcome variable or because the dataset used to train the model is biased, too small, or otherwise unrepresentative of the decision subject population.[53]

Moreover, erroneous decisions can result from a mismatch between the outcome variable and the ideal basis for a decision. A mismatch be-

---

[52] When the outcome variable is a predicted risk or likelihood, one can think of the "true value" as the value that would be computed by a model that took into account every relevant characteristic of the decision subject.

[53] See generally Pedro Domingos, "A Few Useful Things to Know about Machine Learning" (2012) 55:10 Communications ACM 78 at 84–85.

tween the outcome variable proxy and the ideal grounds for a decision can be particularly significant in machine-learning-based decision algorithms because the machine learning process constrains the selection of outcome variables. To train a model that computes reasonably good predictions for an outcome variable, the algorithm designer must have access to a sufficiently large set of data correlating feature values to outcome values. Such data sets are ordinarily available for only a limited selection of outcome variables. The outcome variables for which such data is available may not be close proxies for the decisions' ideal grounds. As a result, algorithm designers may face trade-offs between the accuracy with which the machine learning model proxies for a given outcome variable and the reliability of that outcome variable as a proxy for the ideal bases for decisions.[54]

A "recidivism" algorithm such as the one in our hypothetical, for example, might be intended for use in assessing whether a defendant is likely to commit a violent crime if released on parole.[55] Unfortunately, whether the defendant will actually commit such an offense if released is inherently unknowable at the time of the decision.[56] The next best thing would be to train a model to predict the outcome variable "likelihood that a defendant with particular features will commit a violent crime if released." However, one runs into three types of problems when trying to do this. First, while the relevant outcome variable is likelihood of *committing* a violent crime, one may only have data for rearrest. But rearrest is an

---

[54] See generally Nay & Strandburg, *supra* note 51.

[55] Jessica M Eaglin, "Constructing Recidivism Risk" (2017) 67:1 Emory LJ 59 at 75:

> Selecting the base population for observation is only part of the initial data collection process. To glean information from that base population, developers must specify the outcomes they wish to study and the key variables they wish to observe. This requires developers to translate a problem—here, recidivism—into a formal question about variables. Framing this question requires that developers understand the objectives and requirements of the problem and convert this knowledge into a data problem definition. It is a "necessarily subjective process," requiring developers to finesse a social dilemma such that a computer can automate a responsive answer. Developers frame this question around what they would like to know at sentencing: whether this person will commit a crime in the future. They translate the problem of public safety into a series of questions about the reoccurrence of criminal behaviour and timing. For instance, what events constitute "recidivism"? How far into the future should a tool predict this occurrence? Developers resolve these issues by creating a simple yes–no question for observation in the data set. Yet defining recidivism is less intuitive and more subjective than it may appear. Recidivism means the reoccurrence of criminal behaviour by an individual [footnotes omitted].

[56] See e.g. Julia Dressel & Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism" (2018) 4:1 Science Advances 1.

imperfect proxy for committing a crime. Second, data about rearrest is biased by the fact that only defendants who have been released have had the opportunity to be rearrested and only those offenders (and non-offenders) who are apprehended can be counted. Moreover, predicting violent recidivism presents an additional problem: as rearrests for violent offenses are relatively rare, it is difficult to train a model to predict them because an appropriate sample of data for that outcome variable may be unavailable. Algorithm designers would thus be forced to choose between less accurate, but more meaningful, predictions of violent recidivism and more accurate, but less meaningful, predictions of "rearrest for any offense."

C. Gaming Automated Decision-Making Algorithms

How might such an automated decision-making algorithm be gamed? Though there are three tiers of proxies involved, only the relationships between input data and features and between features and outcome variable predictions provide decision subjects with opportunities for gaming the system. The relationship between outcome variable and true decision criteria is not gameable by decision subjects because decision makers control such relationships.

The decision maker's choice and use of an outcome variable to proxy for the ideal decision criteria is completely beyond the decision subject's control and ordinarily provides no lever for decision subject gaming. Knowing, for example, that our hypothetical algorithm treats "likelihood of arrest" as a proxy for "likelihood of committing a violent crime if released" does not provide a hook for gaming—they cannot do anything to change the fact that decision makers use rearrest as a proxy for recidivism. To take action to obtain a more beneficial outcome, a decision subject must reduce the likelihood of rearrest predicted by the algorithm. To do that it is not enough to know that "likelihood of arrest" is what the algorithm attempts to predict. She also needs information about how likelihood of rearrest is computed by the algorithm.

Because the outcome variable cannot ordinarily be gamed, it follows that it should generally be disclosed, particularly because disclosure of the outcome variable can be extremely valuable for accountability purposes. Disclosing that our hypothetical recidivism algorithm uses likelihood of rearrest as a proxy for likelihood of committing a violent crime would, for example, allow decision subjects and the larger public to debate whether it was appropriate to use it for making parole decisions, in light of its potential inaccuracy and bias.

The proxy relationship between the features and the outcome variable provides more promising, though also limited, opportunities for decision subject gaming. The algorithm's computation is determined entirely by

the decision subject's features and the model's formula for combining them. Decision subjects have no influence over the decision maker's choice of feature set or the algorithm's rule for combining them to estimate likelihood of rearrest. They may, however, be able to modify their behaviour so as to affect their own features. But altering behaviour in this way is not always possible, as we discuss below.[57] Moreover, to game effectively, decision subjects need to know what features the algorithm employs and at least something about how it combines them. In our recidivism hypothetical, for example, prisoners could attempt to improve their chances for parole by displaying good behaviour in prison. But they can only do so effectively if they know that good behaviour "counts", and how much.

In principle, a decision subject could also attempt to game the system by exploiting the proxy relationship between input data and features; in other words, by falsifying input data, rather than changing related behaviour. In many situations where automated algorithms are used, however, this sort of falsification is unlikely because the decision subject has no part in collecting or recording the data and is otherwise unlikely to be able to hack into the relevant databases to falsify them. Specific disclosure of a particular decision subject's input data to that individual or general disclosure of the sources of the input data will often be unlikely to create serious gaming threats and quite likely to be useful for accountability and improving accuracy.

In sum, to game an automated decision-making algorithm, a decision subject must ordinarily be able to alter her features to obtain a beneficial prediction for the outcome variable. While it may sometimes be possible to fake input data, more often than not the only way for a decision subject to obtain a beneficial outcome variable prediction will be to change the behaviour that the features represent.

D. Complexity and the Plausibility of Gaming

Gaming is neither new nor exclusive to machine learning algorithms, but the gaming threat posed by disclosure of such algorithms might arguably be heightened by at least three characteristics of these algorithms: they operate on a large scale, they represent a shift from standards to rules, and they involve a large number of potentially gameable features.[58]

---

[57] See Subparts III.A. & III.B., *below*.

[58] By rules, we mean the use of the term in the legal literature to refer to legal norms that are verifiable ex ante and contrast them with standards, which are verifiable ex post—we do not mean rule-based systems as used in computer science to contrast them with machine learning systems. See Louis Kaplow, "Rules versus Standards: An Economic

The gaming threat is likely to be lower than these arguments would initially suggest, however, because the complex and non-linear ways in which these algorithms combine data tend to make gaming them difficult and costly for decision subjects.

The fact that algorithms operate consistently on a larger scale than any human decision maker is one of their main advantages. However, this same scale makes the implications of gameability more serious. Whereas gaming a gullible human decision maker affects a single decision, one algorithm is likely to replace a large number of human decision makers. If that algorithm becomes susceptible to gaming through disclosure, all those decisions will become gameable, with potentially large social costs.

Moreover, unlike human decision makers, algorithms are rule-like, meaning that they cannot apply judgment to detect gaming and discretion to respond to it.[59] A human decision maker may realize that a decision subject is trying to game the system and discount the unreliable information. For example, a judge or interviewer may notice that decision subjects have begun wearing suits only to look more presentable and therefore discount the meaning of suit-wearing. Algorithms do not have such a capacity to intuitively detect gaming by perceiving such information. Instead, an unmodified model necessarily returns the same result whenever presented with the same input data.

Finally, machine-learning-based algorithms are known for their capacity to base their output predictions on large numbers of features. Because each such feature might be gameable, one might anticipate that algorithms with large numbers of features would be highly gameable. Moreover, the features taken from the "big data" sources that are regularly used by machine-learning-based algorithms often seem to be consumer or social media behaviours that have no obvious strong connections to the outcome variables they are used to predict and are easily modified.

Algorithms with large numbers of features are unlikely to be easily gameable, however, even though it may be easy to find some features that are easy to alter. The key advantage of machine-learning-based algorithms is their ability to make predictions by combining large numbers of features in complex and unanticipated ways. This complexity makes them difficult to game. Complex formulas are hard for decision subjects to understand. And even if a decision subject understands such a complex

---

Analysis" (1992) 42:3 Duke LJ 557 ("the only distinction between rules and standards is the extent to which efforts to give content to the law are undertaken before or after individuals act" at 560).

59 See generally Nay & Strandburg, *supra* note 51.

combination, using that understanding to strategically change the algorithm's outcome is likely to require coordinated modifications to a large number of features. Referring to our earlier discussion, the cost that matters for analyzing the threat of gaming an algorithm is the cost of such a potentially complicated, coordinated change of several features, rather than the cost of changing a single gameable feature.

Taken together, the factors discussed in this Part delineate a set of prerequisites that must be in place for a decision-making algorithm to be gameable. First, the algorithm must use proxies that are loosely connected to the ideal decision criteria. Second, decision subjects must be able to exploit those loose connections by modifying their features. Third, decision subjects must be able to understand how the machine learning algorithm depends on complex combinations of features and to coordinate behaviour modifications to those combinations. The following Part explores when and how this potential may be exploited.

III. When Will Decision Subjects Game

Having explored how proxies determine what aspects of a decision-making algorithm are potentially accessible for gaming, this Part explores when and how decision subjects can exploit those proxies to engage in socially undesirable gaming, as well as what factors make gaming more likely. We begin by introducing some insights from game theory that are useful for the analysis.

A. Signals, Indices, and Gaming Costs

Interactions between decision subjects and decision-making algorithms are, at their core, interactions between decision subjects and the humans that make and apply such algorithms. Crucially, interactions between decision subjects and decision-making algorithms, like more familiar interactions between human subjects and decision makers, are strategic.[60] Insights from game theory, a method that studies strategic interactions among human decision makers, can therefore help to illuminate the algorithmic disclosure question.[61] Signaling models are commonly used to analyze strategic behaviour when there is incomplete information and

---

[60] See generally John von Neumann & Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944). See also Bambauer & Zarsky, *supra* note 2 (establishing "that both the subjects and the designers of algorithms engage in strategic behavior" at 22).

[61] See Spence, *supra* note 7 at 355–57.

have been previously applied to other areas of law.[62] The issues associated with gaming of algorithmic decision-making systems are similar in many respects to those addressed by signaling theory.

Signaling theory addresses situations in which one party must decide how to act considering information that she can observe about another, being aware that the person she is observing can strategically influence what she is able to observe. Here, we imagine situations in which the observing party is a decision maker and the observed party is a decision subject. Signaling theory categorizes the information available to the decision maker into *signals* and *indices*, where signals are pieces of information that the sender (here the decision subject) can influence and indices are pieces of information that are beyond the sender's control.[63] Signals often depend on the subject's behaviour, while indices often depend on the subject's immutable features because one can change one's behaviour (ordinarily at some cost) but, by definition, not one's immutable features.[64] Taken together, *signals* and *indices* that are observable by the decision maker amount to what we have thus far been calling a proxy.

Signals and indices can be either *informative* or *uninformative*, depending on the extent to which they can be relied upon as a basis for a decision. Both indices and signals are uninformative when everyone displays the same signal, so the observing party does not obtain new information from the signal. This often takes place when they are cheap to fake because the relationship between the information they purport to provide and the ideal bases for decision is loose.

To illustrate, consider polygraphs, which purport to detect lying by measuring changes in various physiological characteristics, such as blood pressure, breath rate, and perspiration rate. The theory behind polygraphs is that these physiological signs are *indices* that are not under the control of the subject, and that they are *informative* about whether the subject is telling the truth. As it turns out, lie detectors are considered unreliable proxies for truth-telling for two reasons. First, the physiological changes they measure can result from a variety of emotional and physical states other than lying, whereas some people can lie without ex-

---

62 See Baird, Gertner & Picker, *supra* note 7 at ch 4; McCarty & Meirowitz, *supra* note 7 at ch 8.

63 See Spence, *supra* note 7 at 357–59. Indices are often described in terms of immutable characteristics. For our purposes, the important question is whether a feature can be strategically altered by the decision subject in response to disclosure of the algorithm. Many features that are not immutable in the usual sense are unalterable in this sense.

64 See Bambauer & Zarsky, *supra* note 2 at 15 (introducing a similar definition of immutable proxies, which subjects have *less* ability to change).

hibiting those changes—this is an example of a loose relationship between the polygraph proxy and the characteristic of interest. Second, it is possible to learn to control the physiological symptoms that polygraphs measure, reportedly with comparatively little effort. Thus, it seems that polygraphs, while once introduced as *informative indices,* are probably best described as *uninformative signals*.

To give an example more relevant to algorithmic decision tools, consider an employer who is seeking to promote one of her current employees to a supervisory position. The employer wants to promote someone who will work well in the new position. To try to ascertain which of her employees would do this, she could i) look back at performance reviews; ii) announce that she will soon be promoting a hardworking employee and then observe how hard her employees work; and/or iii) ask the employees whether they would work hard if promoted.

Assume that performance reviews include an assessment of the employees' work habits and are reasonably informative, but are not perfectly informative because some employees would work harder if entrusted with a supervisory position and others would not. Working hard after the announcement and telling the employer that one would work hard if promoted are signals. Saying "I'll work very hard if you promote me" is a classic uninformative signal known as "cheap talk", because it is easy to fake.[65] Working hard after the announcement is more informative because it demonstrates, at a minimum, that the employee is capable of working hard and willing to do so if given incentives. It is more informative than the statement of intent to work hard because it is costlier for the employee to produce. However, it is also not perfectly informative because, though working hard for a few weeks demands a non-trivial investment of effort, an employee might game the system by working hard for just long enough to get the promotion and then slacking off.

Proxies that are tightly connected to desired characteristics are harder to game because of the distribution of signaling costs. If the proxy is equally hard for everyone to acquire, then it will not be informative as a signal because subjects will either acquire it (if it is cheap) or not (if it is costly) independent of their actual characteristics. A proxy that works as an informative signal is cheap to acquire for subjects that have the characteristics that the proxy estimates and costly to acquire for subjects that do not have those characteristics.

---

65 For a minority position, *contra* Scott Alexander, "What is Signaling, Really?" (12 July 2012), online (blog): *Less Wrong* <www.lesswrong.com> [perma.cc/5ZAF-NYZD] (stating that mere assertion of a skill in the job-seeking context may go a long way because society created a system of reputational penalties in which assertions have become credible signals).

A number of issues may increase the cost of gaming for decision subjects in an algorithmic decision process. And the costlier gaming is for decision subjects, the fewer will do it. The likelihood that decision subjects will attempt to game the system depends on the balance between the cost they incur from such efforts and the value they expect to gain from a beneficial decision. Thus, the validity of the algorithm will tend to be maintained when gaming is costly. In the recidivism example, commonly used features, such as drug addiction and postal code, are unlikely to be gamed because they are extremely costly for decision subjects to modify.

The likelihood that disclosing information about a decision-making algorithm will lead to socially undesirable gaming depends on whether, as a practical matter, a decision subject can exploit that disclosure—in other words, whether the disclosure facilitates a gaming strategy that is both feasible and cost-effective. The cost-effectiveness of a gaming strategy depends on whether the value of a better decision outcome outweighs the cost of implementing that strategy. When only limited information is disclosed to decision subjects, they may not be able to ascertain in advance whether a gaming strategy will be successful. In such situations, the value of a better decision must be discounted to reflect the probability that the strategy will succeed.

Putting this together with our earlier analysis, we can see that various proxies that may go into a decision-making algorithm can be distinguished along at least two axes. First, some proxies will not be gameable because they are immutable or prohibitively costly for the decision subject to modify in order to improve the decision outcome (corresponding to the difference between *signals* and *indices*). We will ordinarily think of indices as immutable characteristics of decision subjects, corresponding to immutable features in an automated decision-making algorithm. We note, however, that the choice of outcome variable is ordinarily not gameable for the same reason; it cannot be changed by the decision subject. Second, modifying a proxy to game the system and maintaining the modification until the decision is made will be costly, so that only some of them will be worth gaming by subjects who do not merit a beneficial decision (corresponding to the difference between informative and uninformative signals).

B. Plausible Gaming Strategies

Bambauer and Zarsky's recent article, *The Algorithm Game*, identifies four types of "gaming strategies" that might be used to alter an algorithmic prediction without changing the underlying information that the decision-maker wants to know: avoidance, altered conduct, obfuscation, and

altered input (false reporting).[66] This Subpart builds on the considerations set out above to analyze the conditions under which each of these strategies is likely to be feasible and cost-effective.[67]

A subject engages in avoidance when she takes temporary measures to avoid detection of illegal behaviour by considering information about enforcement activities, such as the location of DUI checkpoints or radar-detector–linked cameras. Avoidance is a feasible gaming strategy when enforcement is sufficiently spotty so that illegal behaviour can be routed around it. It thus reflects looseness in the relationship between a decision proxy and the underlying characteristic of interest.

Bambauer and Zarsky distinguish avoidance from altered conduct, describing it as "avoid[ing] being the subject of an algorithm's model at all."[68] For the purposes of this article, avoidance and altered conduct can be analyzed together. In our view, spotty enforcement is simply a case of an algorithm that employs a loose proxy for the ideal decision-making criterion. Avoidance is feasible when it is possible to locate the "sites" of enforcement and conduct one's illegal activities elsewhere, and cost-effective when it is not too expensive to do so. In essence, these enforcement sites delineate an incomplete feature set used to estimate illegal behaviour. Their sparseness and relatively loose connection to the overall potential for illegal behaviour makes them relatively easy to game when they are disclosed. Outside of law enforcement, avoidance will rarely be a useful strategy because decision subjects often actively seek to benefit from—and therefore voluntarily submit to—the decision-making process. Avoidance is also not a feasible strategy for gaming an algorithm that uses data that is routinely and unavoidably generated and collected as decision subjects go about their daily lives.

Altered conduct refers to action taken to undermine the accuracy of an algorithmic prediction by altering features relied on by the model without changing the underlying characteristics relevant to the decision. It corresponds to one of the two possible levers for gaming that we identified

---

66 Bambauer & Zarsky, *supra* note 2 at 12–13. Bambauer and Zarsky draw readers' attention towards the fact that the strategies, while illustrating the ubiquity of gaming behaviour, are often dependent on a wide range of contexts (*ibid* at 22). In doing so, they raise the possibility that the taxonomy has ambiguous application in some circumstances or may be incomplete (*ibid* at 38). We hope to build on this discussion by exploring further the contexts in which these strategies may or may not be feasible and cost-effective, in an effort to provide further guidance into such ambiguity. One's understanding of the ubiquity of gaming behaviour can be nuanced by the fact that gaming depends on the decision-making context and the nature of the disclosures.

67 See Part II, *above*.

68 Bambauer & Zarsky, *supra* note 2 at 12.

above.[69] Where a positive decision is desirable to decision subjects, as in the employment context,[70] gaming using this strategy aims to turn "true negatives", or accurate detrimental decisions, into "false positives", or inaccurate beneficial decisions. For example, an incompetent employee may seek a promotion thus turning a "true negative" decision into a "false positive" one. Conversely, when a negative decision is desirable to decision subjects, as in the law enforcement context, altered conduct gaming aims to turn "true positives" into "false negatives". For example, if purchasing a plane ticket at the last minute using cash were known to be an important factor in an algorithmic prediction of involvement in illegal drug smuggling, a drug trafficker ("true positive") might decide to purchase a ticket two weeks in advance with a credit card, while still smuggling drugs, to become a "false negative".

Sometimes an altered conduct strategy may not be feasible, for example if the features are immutable (indices). Consider, for example, an algorithm that selects candidates to interview for a job that involves restocking shelves. Because shorter workers need to stand on step ladders to reach the top shelves, the employer prefers taller candidates, all else being equal. Disclosing to a candidate who is 5'1" the fact that the algorithm considers height does not allow her to game the system because her height is an immutable characteristic that she cannot control. Disclosure might, however, provide accountability by allowing short jobseekers to contest the employer's assumptions about the significance of height for job performance, for example by pointing out the advantages that short individuals may have in stocking lower shelves.

Even when an altered conduct strategy is feasible, however, it may not be cost-effective. In the drug trafficker example, purchasing tickets with cash at the last minute presumably benefits drug traffickers in some way—using cash may help them to elude detection by minimizing the data trails left by their illegal activities—while purchasing tickets at the last minute may serve a similar purpose or reflect the way that drug trafficking supply chains function. While it is feasible in principle for drug smugglers to game the algorithm by buying plane tickets further in advance using credit cards, doing so may not be cost-effective if a last-minute cash purchase was a sufficiently valuable "business" technique. As this example illustrates, gaming by altered conduct will be particularly costly for strong proxies which bear a tighter correlation with the ideal decision criteria. In our drug trafficking example, gamers would incur high costs if it was difficult to maintain a successful drug trafficking

---

69 See Subpart II.C., *below*.

70 This is the mirror situation to the parole context, where a "negative" decision, in terms of risk, is desirable for decision subjects.

business while purchasing plane tickets by credit card two weeks in advance.

To obfuscate, our drug ring would send out large numbers of individuals who fit the profile but are not smuggling drugs, along with a few drug smugglers, in the hope that enforcement efforts will be overwhelmed. Effectively, obfuscation seeks to undermine the usefulness of the algorithm by introducing false positives that must be weeded out on a case-by-case basis. It is a method for weakening the proxy relationship between the outcome variable and the ideal decision criteria. Like avoidance, obfuscation has a limited sphere of application outside of law enforcement and it is a special case even in that arena because decision subjects ordinarily do not have the power to control the strength of the proxy relationship between a decision-maker's choice of outcome variable and the ideal decision criteria. For example, it is difficult to imagine how it could be deployed in the employment context.[71] Even where obfuscation is viable in principle, the ability to create false positives depends on having a relatively weak proxy and will usually be too expensive to be cost-effective.

Altered input is defined by Bambauer and Zarsky as falsely reporting information that is used as data to compute a proxy.[72] It corresponds to the proxy relationship between input data and features discussed above.[73] An example of altered input would occur if our drug smuggler were to purchase a plane ticket on the day of travel and then hack into the airline's computer system to make it look as though a ticket was purchased several weeks ago. A more realistic example of altered input would be falsifying one's educational background on a résumé. The altered input strategy has the same goal as the altered conduct strategy—generating a proxy output that improves the decision from the subject's perspective without improving the decision subject's true eligibility for a beneficial outcome. In principle, input data can be altered even when the underlying features are immutable (race or height), costly to change (level of education or experience), or closely tied to relevant underlying characteristics (drug smuggler's airline ticket strategy). As a result, the altered input strategy could, in principle, be viable in circumstances where an altered conduct strategy would not be feasible or cost-effective.

---

71 An example of obfuscation in the private sector would be the following: someone who is taking a test engages in deliberate suspicious behaviour to make the invigilator think they are cheating but, every time she checks on them, they are not doing anything that is dishonest or constitutes cheating. Then, they do in fact cheat as the invigilator is less likely to check on them once more having recalibrated expectations about the results of suspicious behaviour from that person.

72 See *supra* note 2 at 12.

73 See Subpart II.B., *above*.

As a practical matter, however, the altered input strategy is likely to be of limited importance, particularly for the trade-off between disclosure and gaming.[74] While decision subjects can cheaply falsify their résumés or other data collected directly from them, this strategy is undoubtedly anticipated by decision makers who can often thwart it by relying primarily on third-party or verified data sources. Moreover, requesting information from decision subjects implicitly discloses the fact that the requested information is likely to be used in the decision process. In those cases, disclosure may not heighten the incentive to fake the information significantly. When data is obtained from third-party sources, it is usually more difficult and costly for decision subjects to carry out an altered input strategy. Hacking into and altering third party data sources requires specialized skills and is likely to be too costly and difficult for most decision subjects even if the fact that an algorithm employs such features is disclosed.

The effects of particular disclosures on the feasibility and cost-effectiveness of the gaming strategies identified by Bambauer and Zarsky will depend on the decision-making context and the nature of the disclosures.[75] In terms of the above analysis of the three layers of proxies involved in automated decision-making algorithms,[76] altered input falsifies the relationship between input data and features, obfuscation affects the strength of the relationship between outcome variable and ideal decision criteria, and avoidance and altered conduct target the relationships between features and outcome variables.

Disclosure of information about secret automated decision-making algorithms will increase the threat of gaming by obfuscation or altered input only in a narrow range of circumstances. If a decision maker asks a decision subject for data, rather than collecting the information from a third party, such data can perhaps be falsified independently of whether

---

74 Bambauer & Zarsky, *supra* note 2, also suggest limitations that may arise in the context of altered input that suggest its limited importance. For instance, they allude to limitations relating to the prevalence (or lack thereof) of altered input strategies on the basis of some people's disposition or personal moral code, e.g. personal traits and religious convictions (*ibid* at 29). Another limitation they offer is the common prohibitions on altered inputs (*ibid* at 43). In these situations, gaming is risky because of the risk of liability exposure. One could infer from this consideration on risk the suggestion that altered input may not pose the same concerns for gaming as other strategies.

75 See Bambauer & Zarsky, *supra* note 2 (recognizing that gaming strategies are influenced by other variables beyond our proposed decision-making context and the nature of the disclosures, for example, suggesting that "even less blunt forms of gaming ... will be off limits to some individuals who understand that this conduct will have negative effects on the algorithm designer or on other subjects" at 29–30).

76 See Subpart II.B., *above*.

there is disclosure. Falsified input could also be a concern when this information is not requested and the decision subject still has the ability to fake it, but this will be infrequent. The most likely threat from disclosing information about a secret automated decision-making system is an increase in gaming by altered conduct—including avoidance, which we consider a type of altered conduct in the enforcement context.

In sum, we believe altered conduct is the main concern for gaming, where we view avoidance as a special case of altered conduct. Obfuscation and altered input will be unfeasible or overly costly in most situations of practical interest. In terms of the three tiers of proxy relationships we identified,[77] the altered input strategy affects the proxy relationship between input data and features, while obfuscation affects the relationship between outcome variables and ideal decision criteria. Altered conduct uniquely targets the relationships between features and outcome variables. Our analysis thus focuses on altered conduct, which is the gaming strategy most likely to pose a realistic threat to algorithmic decision making.

C. Socially Desirable Strategic Responses to Disclosure

The *signal/index* and *informative/uninformative* distinctions from signaling theory can generally be used to identify features that are likely to be gamed if they are disclosed. Those distinctions do not, however, account for the possibility that a decision subject may strategically invest in a feature modification that changes her underlying eligibility for more beneficial treatment. Not every strategic attempt by a decision subject to obtain a more beneficial outcome is socially undesirable gaming. To illustrate the point, consider the following example: someone wanting to obtain a loan pays her debts on time and keeps a high salary-to-borrowing ratio after learning that these factors are important in maintaining a high credit score, which lenders use a proxy for likelihood of loan repayment. The potential loan applicant's altered conduct, while strategic, hardly can be described as "gaming the system." The disclosure of the factors that improve a credit score served as a tutorial in good financial behaviour. Though her behavioural changes may have been strategically motivated by her desire to eventually obtain a loan, it seems likely that her successful adoption of those behaviours will improve her true creditworthiness.[78]

---

77 See *ibid*.

78 This hypothetical assumes that high credit scores are relatively closely tied to good creditworthiness. Roughly, this means assuming that the false positive rate is low. That assumption says nothing about the much more contentious question of whether there are significant numbers of creditworthy individuals with low credit scores (rough-

As another example, consider a recidivism risk assessment that relies on historical variables, such as number and timing of prior arrests and convictions,[79] or, as in our hypothetical, gang membership or prior employment. It is already a stretch to imagine that disclosure would motivate a potential defendant to modify these sorts of behaviours in advance so as to obtain a more favourable recidivism prediction after some future arrest. But suppose an individual did strategically respond to disclosure of a recidivism algorithm by taking steps to reduce arrests and convictions, avoid gangs or obtain employment. It seems entirely likely that taking those steps would reduce the individual's "true" recidivism risk and thus be both individually and socially beneficial.

In other words, while we discussed examples of people gaming a signal to get a better outcome, one can also think of examples of people attempting to improve their chances of a beneficial decision by changing the underlying characteristic that the signal portrays. The result of such efforts is different than that of someone gaming the system. The difference is not in the amount of investment made compared to people gaming the system, but in the social implications of that investment.

A decision subject's strategic behaviour might also compensate for errors in the proxy relationship between the outcome variable and the ideal decision criteria. Because the "big data" required for machine learning constrains the selection of outcome variables, algorithm designers must often choose outcome variables that are imperfect or noisy (or even biased) proxies for the ideal decision-making criteria. That noise (or bias) means that decisions based on the algorithm's predictions will sometimes be mistaken even when the algorithm correctly predicts the outcome variable.[80]

---

ly, the false negative rate). A counter-example of when credit status may be a loose proxy, and therefore susceptible to gaming, would be when credit status is used as a proxy in the hiring process. Indeed, employers frequently use credit reports as screens for potential hires. Given that there is little empirical data to suggest that credit status is a strong predictor of employee productivity, credit scores are arguably a loose proxy in this context. See e.g. Andrew Weaver, "Is Credit Status a Good Signal of Productivity?" (2015) 68:4 IRL Rev 742 at 743–44.

79 See *ibid* at 81–82.

80 For an analysis of how noisy decision making by humans may introduce quasi-experimental variation into training datasets which may benefit machine learning, see Bo Cowgill, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening" (2018), draft available online (pdf): *Institute of Labor Economics* <conference.iza.org> [perma.cc/Q45A-UCTZ] (arguing that noisy, biased, human decision makers working with datasets create experimental variation in the training data that facilitates de-biasing, rather than codification of pre-existing bias. Without noise, according to the author, new learning technology has no information about counterfactual decisions and their outcomes).

For example, returning to our recidivism hypothetical, an individual who is not likely to commit a violent crime if released pending trial, might be likely to be rearrested either because she is likely to commit a less serious crime, or because characteristics such as where she lives, her race, or her social group make her likely to be rearrested even if she commits no crime at all. She may well be mistakenly detained simply because the algorithm employs an inaccurate outcome variable proxy for likelihood of violent crime. Now suppose such a decision subject strategically alters or fakes her features so that the algorithm makes a more beneficial (but wrong) prediction of her likelihood of rearrest. For example, she might lie about a drug addiction so that the algorithm will deem her unlikely to be rearrested when she is, in fact, likely to be rearrested for a non-violent drug offense. If the policy goal is to grant parole as long as a prisoner is unlikely to be violent, this strategic behaviour fools the algorithm, but produces a socially desirable result; she has, in effect, corrected an error created by using an outcome variable that is a bad proxy for the ideal decision criteria.[81]

One may or may not sympathize with the defendant's strategic behaviour in the above example, but the potential for strategic error correction comes into sharper relief if one considers strategies of applicants for employment or housing involving altering names or other input data to avoid race or gender bias introduced by training data that reflects the biases of previous hiring or rental decisions. When a proxy is systematically awry for some sub-group, disclosure can facilitate strategies that systematically lead to socially beneficial error correction, rather than to socially undesirable gaming.

In sum, individuals may respond to disclosure of information about decision-making algorithms by making behavioural changes that qualify them for more beneficial outcomes. Those changes, even if strategic, and even if made purely in response to the disclosure, do not constitute socially undesirable gaming. Perhaps more controversially, decision subjects may sometimes provide socially beneficial error correction even by strategically "fooling" a decision-making algorithm when the algorithm's outcome variable is an imperfect proxy for the ideal decision criteria.

---

[81] While Bambauer and Zarsky believe that gaming will usually increase error, they also raise the possibility that gaming could improve accuracy. See Bambauer & Zarsky, *supra* note 2 ("[i]n rare cases, gaming could improve accuracy if the conduct of gaming does not change the key characteristic of the gamer in any way, but the gaming itself helps ambitious, creative, or attentive subjects distinguish themselves to correct for preexisting errors that would have otherwise been biased against them" at 25 n 103).

D. Conditions for Socially Undesirable Gaming

Ultimately, the overall performance of a decision-making algorithm will be determined by both the noisiness of the proxies employed and the extent to which the algorithm is gamed. These factors are not independent. As discussed above, strongly tied proxies are generally difficult to game because there is usually some underlying reason for the strength of the relationship. However, even loose proxy relationships can be relatively impervious to gaming.

First, disclosing proxy relationships that cannot be exploited by decision subjects creates no potential for gaming. Such ungameable proxies ordinarily include outcome variables, input data, and decision subject features that are effectively unalterable (indices).[82]

Second, even when a feature is alterable in principle, there will be no gaming unless decision subjects decide that investing in altering that feature is cost-effective in light of the benefits of an improved decision outcome.

Third, gaming an automated algorithm will often require coordinated alternations of a number of features, which may not be cost-effective in combination.

Fourth, some features are of such obvious relevance to decisions that decision subjects will not need to be told that decision makers are likely to take them into account. For such obviously relevant features, only detailed disclosure about how the feature plays into the decision can lead to increased gaming.

Fifth, even if disclosure motivates a decision subject to invest in strategically modifying a feature, the result may not be socially undesirable gaming if modifying the feature improves the subject's true eligibility for a positive outcome. Developing good financial habits in hopes of getting a loan, studying hard to get a good GPA so as to get a good job, and similar alterations leave the decision subject more deserving of a positive outcome than she was before. Similarly, when disclosure does allow decision subjects to strategically fool the algorithm, the result will sometimes be socially beneficial error correction, rather than socially detrimental gaming. In particular, strategic behaviour by decision subjects can be socially beneficial if an algorithm explicitly or implicitly employs socially undesir-

---

[82] This may be used to differentiate using race, for example, as an index in machine learning versus using a signal for race as a feature. Race is not gameable (setting aside the question of "passing") whereas some proxies will be gameable. When the feature is unobservable, we may use indices or proxies for it, but sometime proxies are used because the indices in themselves are also unobservable.

able stereotypes based on race, gender, age, or other protected characteristics as proxies for decision criteria.[83]

E. Structuring Nuanced Disclosure Regimes

Disclosure about decision-making algorithms is not an either-or matter and can have a variety of different policy goals.[84] As Barocas and Selbst point out in a recent article addressing the issue of explanation, useful disclosure of information about a machine learning algorithm could take many forms.[85] The sorts of disclosure that are feasible and desirable for policy reasons in a given situation may or may not provide information that a decision subject could exploit effectively to game the system.

Because effective gaming requires fairly extensive information about the features employed and how they are combined by the algorithm, it will often be possible for decision makers to disclose significant information about what features are used and how they are combined without creating a significant gaming threat. Here we have given a few simple illustrations of disclosures that would promote accountability without facilitating gaming, but many sorts of disclosure regimes are possible. Banks and credit bureaus seem to understand this idea, for example. Consequently, they disclose the features used to make decisions without disclosing the exact weights. Canadian banks, for example, have open mortgage risk-assessment metrics post-2008. We do not see gaming for mortgage

---

83 See Edmund S Phelps, "The Statistical Theory of Racism and Sexism" (1972) 62 Am Econ Rev 659 at 659. See also Kenneth J Arrow, "The Theory of Discrimination" in Orley Ashenfelter & Albert Rees, eds, *Discrimination in Labor Markets* (Princeton: Princeton University Press, 1973) 3. See also Shelly Lundberg & Richard Startz, "On the Persistence of Racial Inequality" (1998) 16:2 J Labor Econ 292; Fang Hamming & Andrea Moro, "Theories of Statistical Discrimination and Affirmative Action: A Survey" in Jess Benhabib, Alberto Bisin & Matthew O Jackson, eds, *Handbook of Social Economics*, vol 1 (Radarweg: Elsevier, 2011) 133 at 134–35.

84 To illustrate the variety of policy goals, consider the following. The Center for Data Innovation, on the one hand, lists their policy goals for algorithm accountability as seeking to minimizing the risk of harm and negligence caused by using algorithms, see Joshua New & Daniel Castro, "How Policymakers Can Foster Algorithmic Accountability" (2018) at 29, online (pdf): *Center for Data Innovation* <datainnovation.org> [perma.cc/Y4VL-JVSZ]. On the other hand, the Institute of Electrical and Electronics Engineers (IEEE) call for disclosure as a way to increase public trust and reliability in machine learning: see Katyal, *supra* note 1 at 110, citing Institute of Electrical and Electronics Engineers, "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems" (2016) at 44, online (pdf): *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* <standards.ieee.org> [perma.cc/L92A-LCDF].

85 See generally Andrew Selbst & Solon Barocas, "The Intuitive Appeal of Explainable Machines" (2018) 87 Fordham L Rev 1085.

metrics even when disclosed because they are features that cannot be gamed: income, employment, liquid assets, etc.

For this reason, cavalier assertions that disclosure of machine-learning-based algorithms will facilitate "gaming" are highly contestable, particularly when one considers the variety of possible disclosures that could be made. For example, suppose the list of features is disclosed. Even if some or all those features can be cheaply altered by decision subjects, gaming will be difficult for decision subjects as long as they do not know how those features combine to determine the outcome variable. This problem arises even when the predictive model is linear, so long as the weights given to each feature are not disclosed. The difference between disclosing a list of features and disclosing how an algorithm combines them is even more stark for the sorts of machine-learning-based models that are not fully understandable even to decision makers who design and employ them.[86]

Identifying not only whether gaming is a plausible concern, but also what particular information facilitates gaming, allows, at a minimum, for the selective disclosure of other types of information. Oftentimes, the information that will be most useful for accountability is the features of the model, given that features tend to be more easily understandable than the formulas for combining them. For example, decision makers can disclose the data that is used to assemble the features, or the features that are used to form the outcome variable, without disclosing their weights or the importance that is given to any of the data and features in the overall assessment. This type of disclosure will not always be possible, for example for deep learning algorithms that process raw, unlabeled data.[87] But these models are rarely used in the sorts of individual decision-making contexts we have been discussing. And even with these models, some disclosure is always possible; particularly, information about their training data can be disclosed. If privacy or other concerns prevent disclosure of the training datasets, the provenance of the data can be disclosed instead. Disclosing training data sources alone is oftentimes helpful for accountability, for example, to evaluate whether the learned model is biased towards a group of individuals.

---

[86] See Informatics Europe & European Union Association for Computing Machinery (EUACM), "When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making" (2018) at 9–10, online (pdf): *Informatics Europe* <www.informatics-europe.org> [perma.cc/G529-PFEX] (explains why, for many machine learning models, the decision makers who employ them are not able to explain the models themselves).

[87] See Goodfellow, Bengio & Courville, *supra* note 4.

These nuances are of great importance to real world applications, yet they are too often elided or ignored in policy discussions of the potential for "gaming the system." Often it is simply asserted that once an algorithm is disclosed in some unspecified manner, it can and will be gamed to socially undesirable effect.[88] Evaluating the specific ways in which algorithmic disclosure does or does not facilitate socially undesirable gaming is a contextual task, which we do not tackle here. The framework presented here is intended to help us begin to think about these specifics. If a disclosure does not teach decision subjects how to improve their chances of favourable treatment by taking some feasible action, it has not made the algorithm gameable. To game, subjects need a level and type of disclosure that gives them enough information about proxy composition to determine how to modify the proxy in a favourable direction. In many situations, even the most fulsome disclosure will not create a serious threat of socially undesirable gaming. However, even when complete disclosure would create such a threat, it is possible to use limited disclosures to advance important policy goals.

IV. When Will Decision Makers Game

We explored above how interactions between decision subjects and people developing and applying decision-making algorithms are strategic, inquiring about what makes specific algorithms gameable,[89] and when decision subjects are likely to game them.[90] But there is an added layer to such analysis. While decision subjects seek better outcomes for themselves, algorithm creators and users pursue their own private ends. None of these players can be assumed to be acting in society's interest. Decision makers' choices about whether and what to disclose about the decision-making algorithms they employ are also strategic and need not be socially optimal, as are their choices about whether to invest in creating less gameable algorithms. As a result, one should closely examine decision makers' assertions about the threat of gaming because (i) they may have

---

88 See e.g. Rayid Ghani, "You Say You Want Transparency and Interpretability?" (29 April 2016), online (blog): *Rayid Ghani* <www.rayidghani.com> [perma.cc/8TYR-KPXP]; Tarleton Gillespie, "Algorithmically Recognizable: Santorum's Google Problem, and Google's Santorum Problem" (2017) Information Communication & Society 20:1 at 67–68; Michael Veale, "Logics and Practices of Transparency and Opacity in Real-World Applications of Public Sector Machine Learning" (2018) at 3, online (pdf): *Cornell University* <arxiv.org> [perma.cc/X6Z4-4KYZ]. See also Kartik Hosanagar & Vivian Jair, "We Need Transparency in Algorithms, But Too Much Can Backfire" (2018), online: *Harvard Business Review* <www.hbr.org> [perma.cc/59J8-8HNL].

89 See Part II, *above*.

90 See Part III.

socially suboptimal motives for opacity or (ii) they may have means other than secrecy to minimize gaming.

A. Private Welfare and Social Welfare

Decision makers' choices about disclosure can be misaligned with social welfare for two reasons. First, decision makers may not adequately account for the value of disclosure to decision subjects and society at large.[91] Second, decision makers may fail to account for the social costs of inaccuracy and bias.[92]

While secrecy can sometimes prevent decision subjects from gaming the system, it can also mask socially undesirable algorithm design. Thus, a fraction of the decision makers who argue that secrecy is necessary to avoid undesirable gaming may be making such claim strategically. Trade secrecy claims, for example, while ostensibly based on worries about free riding competitors, [93] can be used strategically to avoid accountability. As a result, arguments for secrecy based on a threat of gaming by decision subjects should be evaluated carefully, especially when made by government officials or by decision makers subject to anti-discrimination laws or consumer protection regulations.

Disclosure of the proxies and procedures used in decision making often has the potential to confer significant social benefits by promoting accountability, improving decision accuracy, deterring or exposing bias, arbitrariness, and unfairness, permitting decision subjects to challenge the

---

[91] See e.g. Citron and Pasquale, *supra* note 29; Pasquale, *supra* note 44; Selbst & Barocas, *supra* note 85 at 1118–19.

[92] There can be different types of bias in computer systems. Batya Friedman and Helen Nissenbaum identify three types of bias: personal biases of decision makers, technical bias, and bias that emerges after the design has been completed. See Batya Friedman & Helen Nissenbaum, "Bias in Computer Systems" (1996) 14:3 ACM Transactions on Information Systems 330 at 333–36.

[93] Indeed, strategic assertions of secrecy to avoid accountability may begin with specialized vendors, who often cloak their algorithmic tools with secrecy, potentially avoiding accountability not only to decision subjects, but also to the ultimate decision makers who rely on them. We mostly ignore these complications here to focus on strategic interactions between algorithm users/designers and decision subjects. We note as an aside, however, that claims that trade secrecy is needed to deter free-riding competitors and incentivize innovation may be dubious or even pretextual when network effects or other first-mover advantages are significant. See e.g. Yafit Lev-Aretz & Katherine J Strandburg, "Better Together: Privacy Regulation and Innovation Policy" 22 Yale JL & Tech (forthcoming 2020); Eli Siems, Nicholas Vincent, & Katherine J Strandburg, "Trade Secrets and Markets for Evidential Technology" [unpublished].

factual or other bases for erroneous decisions, and to undertake the socially beneficial strategic behaviours discussed above.[94]

If decision makers could be trusted to have society's interests at heart, they would weigh these benefits of disclosure against the potential costs of socially undesirable gaming and make socially optimal decisions about whether, and in what detail, decision-making algorithms should be disclosed. However, decision subjects are not the only ones who can game.[95] Many of the social benefits of disclosure arise precisely because disclosure addresses conflicts between decision-maker incentives and the public good. The threat of gaming can itself be wielded strategically by decision makers seeking to avoid accountability, cut corners, cover up bias, or otherwise place their own interests above those of society at large.

B. Decision Makers as Imperfect Agents of Society's Interests

Moreover, decision makers may have self-serving and strategic incentives to hide the details of their decision-making bases and procedures from those to whom they are accountable, such as supervisors, government officials, or the public at large.

The fact that decision makers are imperfect agents of society's interests is hardly news. There is a large literature associated with the problem of "public choice": the ways in which the private interests of government actors can distort their behaviour away from the public interest they have been appointed to serve.[96] Explanation of government decision

---

94 See Subparts III.C. to III.E., *below*.

95 See Zarsky & Bambauer, *supra* note 2 at 3, stating that algorithmic decision makers game the system by responding to decision subjects gaming by changing their behaviour as a means of discouraging gaming or reducing the effects of gaming. As per Zarksy & Bambauer: "[w]ithin limits, people game the system for a range of altruistic and self-serving reasons. And algorithm designers game right back, using countermoves to discourage gaming or to reduce its effects" (*ibid*).

96 See e.g. Jonathan R Nash, "Economic Efficiency versus Public Choice: The Case of Property Rights in Road Traffic Management" (2008) 49:3 BC L Rev 673 at 681, explaining public choice as follows:

> In general, public choice theory looks at government action as the result of a "market for government action". Under this model, government actors take steps that are designed to maximize their chances of remaining in power. For legislators, this means taking actions that maximize their re-election chances.
>
> The public choice model predicts that government actors will act in response to pressure brought by interest groups. Interest groups give rise to demand for certain government actions, and government actors offer supply in the form of support for different government actions. Thus, an action is more likely to be taken when it is (i) demanded by more, and more powerful, in-

making is a core requirement of due process (procedural fairness) that is intended, at least in part, as an accountability mechanism. Consequential private sector decisions are also subject to legal disclosure requirements. For example, fair credit laws demand a certain level of disclosure to applicants about the bases for loan denials.[97] In other arenas, such as employment and housing, while the law does not require disclosure of decision-making criteria, it does prohibit reliance on certain characteristics, such as race, gender, age, and disability.[98]

The principal-agent problem in government decision making manifests in various ways. Government decision makers may, for example, invest less in decision making than would be socially optimal, or they may over-emphasize certain kinds of mistakes and under-emphasize others, as when an elected judge places undue emphasis on the reputational risk associated with releasing defendants compared to the social and individual costs of unnecessary detention. Private decision makers serve less obviously as agents of the public interest, but in contexts such as employment, education, housing, and credit, their decisions about issues such as how much care to take to avoid bias and discrimination may also have significant externalities affecting the public interest.

Automating some or all the decision-making process is not a silver bullet for avoiding such principal-agent problems;[99] it simply moves them upstream to the point at which the automated process is designed or procured. Algorithms are not autonomous; algorithms are built, trained, and implemented by people.[100] People must select what data are useful to es-

---

terest groups, and (ii) supported by more, and more powerful, government actors. In the environmental arena, relevant interest groups are likely to be industry actors or groups, and environmental interest organizations.

97 See e.g. *Equal Credit Opportunity Act*, 15 USC § 1691(d)(2) (2012); *Dodd-Frank Wall Street Reform and Consumer Protection Act*, Pub L No 111-203, 124 Stat 1376 at § 1474 (codified as amended at 12 USC § 5301–5641 (2012)).

98 See e.g. *Canadian Human Rights Act*, RSC 1985, c H-6, s 7.

99 See generally Archon Fung et al, "The Political Economy of Transparency: What Makes Disclosure Policies Effective?" (2004) Ash Institute for Democratic Governance and Innovation, John F Kennedy School of Government, Harvard University Working Paper OP-03-04, online: *Harvard University* <ash.harvard.edu> [perma.cc/7H9X-N3EX] ("[u]sers of transparency systems have diverse interests. They may include consumers, voters, employees, suburbanites, inner city residents, competitors, organizations representing businesses or consumer interests, legislators, government agencies, and regulators themselves. They may be casually or intensely interested in new information. Their goals may or may not coincide with those of policy makers" at 10).

100 See Jack M Balkin, "The Three Laws of Robotics in the Age of Big Data" (2017) 78:5 Ohio State LJ 1217 at 1223; Zachary C Lipton, "The Mythos of Model Interpretability" (2018) 16:3 Queue - Machine Learning 1 at 3; Ignacio N Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:2 Hastings LJ at 6–20; Nick Seaver,

timate features, what features are important enough for the algorithm to consider in order to determine the output variables, and what output variables to use as an estimator for the ideal decision.

The tension between society's interests and decision makers' personal interests not only affects the way decisions are made, but also gives decision makers incentives to avoid accountability and embrace secrecy. When decision makers value the private benefits afforded by secrecy, they can be expected to exaggerate the threat that disclosure will enable decision subjects to game the system and thus degrade decision-making performance.[101] This is not to suggest that decision makers are unconcerned with making sound decisions or that their warnings about the potential for gaming should go unheeded. The point is only that, when push comes to shove, decision makers may not make socially optimal trade-offs between investments in accuracy and the social costs of various sorts of errors and may exaggerate the threat of gaming in order to protect their private interests.

C. Strengthening the Proxy to Avoid Gaming

As discussed earlier, there are various levels of proxies involved in decision-making algorithms. Legal rules are sometimes framed explicitly in terms of indirect proxies for policy targets.[102] Weak proxies, in turn, are more likely to be gameable than are strong proxies.[103] As a result, weak proxies and gaming will often go hand in hand. Decision makers (or, more to the point, algorithm providers) can respond to the association between weak proxies and gaming by hiding the fact that their decision procedures are not based on robust proxies. In those cases, the threat of gaming, which will often be real for algorithms employing weak proxies, provides a convenient excuse for such secrecy.

"What Should an Anthropology of Algorithms Do?" (2018) 33:3 Cultural Anthropology 375 at 378.

[101] See generally Lee Rainie & Janna Anderson, "Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight" in Lee Rainie & Janna Anderson, eds, *Code-Dependent: Pros and Cons of the Algorithm Age* (Washington, DC: Pew Research Center, 2017), online: *Pew Research Center* <www.pewinternet.org> [perma.cc/S9EC-AH83] (canvassing academics, technology experts and industry insiders who all called for algorithmic transparency. For example, David Lankes at University of South Carolina pointed this out saying, "unless there is an increased effort to make true information literacy a part of basic education, there will be a class of people who can use algorithms and a class used by algorithms").

[102] See Bambauer & Zarsky, *supra* note 2 ("[b]ut law is also replete with examples in which gaming is either directly supported or frustrated by design and intention" at 33).

[103] See Subpart II.A., *above*.

However, secrecy is not the only available way for decision makers to discourage gaming. Instead, decision makers can respond to the threat of gaming by devising stronger proxies, thus simultaneously improving decision performance and making gaming more difficult. By adopting and disclosing more accurate proxies, decision makers can sometimes encourage decision subjects to invest in developing features that improve their qualifications for positive decision outcomes, often simultaneously producing better results for decision makers. Suppose a software company had been screening potential employees by looking at data about subscriptions to the top PC and Mac magazines and websites. This proxy is likely to be easily gameable, in part because it is a weak proxy for software engineering skills. If the company starts basing its screening on college grades in software engineering classes instead—and if college grades are a reasonably sound proxy for performance—the company can benefit from both adopting and disclosing its new criterion.

Gaming a strong proxy tends to be costly for decision subjects. The only way to game the good grades outcome variable (a proxy, in turn, for academic performance) without putting in the hard work of studying and learning the material is to falsify input: to fake the grades. However, grades are easily verifiable by requesting transcripts from the college or university. Hacking into the university system to fake one's transcript is likely to be a risky and costly strategy (even for software engineers), and thus rarely cost-effective. If anything, disclosing the good grades criterion is likely to benefit the employer by incentivizing more potential employees to invest in obtaining good grades and, presumably, learning software engineering skills.

The threat of gaming, coupled with a disclosure requirement, can motivate decision makers to devise and adopt more accurate proxies for the idea decision-making criteria. Unless upgrading the proxy is too costly, this strategic response can be beneficial for decision subjects, decision makers and society overall. Secrecy is not the only response to gaming. Decision makers can often change the proxy instead.

Conclusion

To decide whether mandating disclosure of information about a decision-making algorithm will undermine its effectiveness, policy-makers should begin by asking whether the algorithm and the proxies it employs meet the basic prerequisites for socially undesirable gaming. Sometimes it will be clear from this initial inquiry that at least some aspects of the proxy or algorithm can be disclosed without creating serious gaming issues.

The concepts discussed here create a framework for analyzing when policy-makers need not be concerned with gaming when deciding whether

to mandate disclosure of decision-making algorithms and what disclosure to require. We showed that disclosure cannot seriously increase the threat of socially undesirable gaming unless several prerequisites are met:

i) Decision-making proxies are weak (loosely tied to the ideal decision-making criteria), so that there is enough room for gaming;

ii) The proposed disclosure must pertain to features that are sufficiently modifiable by decision subjects;

iii) Modifying those features must be cost-effective, individually and in combination;

iv) Modifying those features must improve the proxy without improving the decision subject's true eligibility for a beneficial decision.

If a proposed disclosure requirement does not create all of these prerequisites for socially undesirable gaming, secrecy arguments based on the threat of gaming should be discounted. This is, of course, only a sufficient condition for disclosure, but we think it will often be enough to undermine secrecy arguments premised on gaming for particular algorithms and to inform tailored disclosure requirements.

Even if a proposed disclosure meets these prerequisites and would facilitate some gaming, however, secrecy may not be socially optimal if the social benefits of disclosure outweigh the social costs of allowing some gaming or implementing a more gaming-resistant proxy.[104] All in all, societal losses from decision subject gaming and society's potential benefits must be evaluated in light of the specific disclosures to be made.

---

[104] While this question is beyond the scope of the present article, it is the topic of a work in progress that builds on this article.