Transparency in Automated Decision-Making and Profiling
Justifications, Limits, and New Directions for Canadian Privacy Policy

Michael Gora

University of Ottawa
August 2018

Abstract
Canadian lawmakers may soon find themselves looking to amend PIPEDA to regulate the private sector's use of automated decision-making and profiling systems. If such an amendment were to be made, it must not rest upon the sole assumption that increasing transparency in the form of expanded access to information will land individuals in a better position to understand or challenge how their personal information is used. This paper presents three reasons why transparency alone is incapable of achieving these aims. First, simply having access to information does not imply that it is capable of being understood or used to generate accountability. Second, it's not always clear that transparency trumps secrecy. Rather, a fully transparent automated system may conflict with an organization's desire to prevent gaming, to safeguard confidential information, and the public's reluctance to grant others' transparency at the expense of their own privacy. Finally, certain characteristics of machine-learning techniques which often underlie automated decision-making and profiling systems may make it undesirable and impossible to generate information that is capable of being understood or used to facilitate accountability. To overcome these limits, I suggest Canadian lawmakers take on board the following two recommendations. First, organizations using personal information for automated decision making or profiling must grant individuals access to meaningful information defined as that which permits him or her to understand or challenge an automated decision or instance of profiling having due regard for variability in age, technical and legal literacy, time, and knowledge about avenues of institutional recourse. This means utilizing audio-visual tools and data-visualization techniques to generate information at multiple levels of technical and legal complexity such that it would be considered subjectively and instrumentally useful to an epistemically diverse audience. Second, where organizations are unwilling (on account of secrecy-promoting concerns) or unable (on account of the characteristics of machine-learning) to provide individuals with this meaningful information, they must have their automated decision-making and profiling systems pre-emptively verified by a certification body within the Office of the Privacy Commissioner. Similar to the GDPR, these regulatory recommendations ought to apply to solely automated decision-making or profiling systems that produce legal effects or similarly significantly affect the individual whose personal information is used.

Table of contents

Introduction

The *Personal Information Protection and Electronic Documents Act* (*PIPEDA*) gives individuals the right to access information about how organizations use their personal information.[1] However, when it comes to accessing information about how organizations use personal information to feed their automated decision-making or profiling systems, Canadian legal academics have flagged that *PIPEDA* is falling behind.[2] What is it about automated decision-making and profiling *vis-a-vis* old-fashioned human decision-making and profiling that accounts for this regulatory gap? It appears to be that whereas human decision-makers are at least in principle always able to generate information about how an individual's personal information was used to profile or make a decision about him or her,[3] automated decision-making and profiling systems will not generate such information unless they are explicitly programmed to do so.[4] Without access to this information, individuals are less able to understand and control how their personal information is being used and less able to hold an organization accountable when it appears to have been used unlawfully. If this interpretation is correct, should *PIPEDA* be amended to ensure that individuals are able to access meaningful information about how their personal information is used in the process of automated decision-making or profiling?

---

[1] *Personal Information Protection and Electronic Documents Act* (*PIPEDA*), SC 2000, C 5, Principle 4.8 Openness and Principle 4.9 Individual Access.

[2] Standing Committee on Access to Information, Privacy and Ethics, Number 053, 1st Session, 42nd Parliament (Thursday, March 23, 2107) Tamir Israel at 17:16, online: https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/meeting-53/evidence; Standing Committee on Access to Information, Privacy and Ethics, Number 054, 1st Session, 42nd Parliament (Tuesday, April 4, 2107) Ian Kerr at 16:32, online: https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/meeting-54/evidence; A Local Law (New York City) in relation to automated decision systems used by agencies online: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

[3] Even if it's a *post-hoc* explanation or a lie, the certain availability of that information is the only way that it will reveal itself as such.

[4] Finale Doshi-Velez et al, "Accountability of AI Under the Law: The Role of Explanation" (2017) Berkman Klein Center for Internet & Society online: https://dash.harvard.edu/handle/1/34372584?show=full.

Such a move would have Canada fall in line with the European Union's General Data Protection Regulation (GDPR) which despite being hotly debated is unanimously considered to provide individuals with some form of meaningful information about the logic involved in an automated decision-making or profiling system.[5] Aside from that, amending *PIPEDA* in this way would bolster transparency, which upon first glance is a move that appears incapable of objection. Most often spoken about in the context of public governance,[6] transparency is a somewhat amorphous concept to pin down.[7] For the purposes of this discussion though, I conceive of the relationship between transparency and automated decision-making and profiling in the following way. First of all, transparency instinctively denotes a sense of being able to see into or through something. But given that an argument is being made for wanting automated decision-making and profiling systems to be transparent (rather than opaque), it's implied there is a purpose or a reason for wanting to see into or through an automated decision, as if there was something about it that we want to understand.[8] In the context of administrative decision-making, Patrick Birkinshaw suggests the information we seek to understand falls along two axes. We're interested in a decision-maker's *rationale* – the "procedures, information, reasons, and the facts on which the reasons are based" – as well as the *process* that led it there – "as in the deliberations and negotiations of the

[5] Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Advancement of Such Data, and repealing Directive 95/46/EC, 2016 O.J. L 119/1 [hereinafter the General Data Protection Regulations or GDPR].

[6] As a matter of administrative law, the Supreme Court of Canada ruled in *Baker v Canada (Minister of Citizenship and Immigration)* [1999] 2 SCR 817 at para 44 that the principle of procedural fairness requires that individuals are "entitled to fair procedures and open decision-making." Moreover, in situations "where the decision has important significance for the individual, when there is a statutory right of appeal, or in other circumstances" the majority held that it maybe incumbent upon decision-makers to provide written explanations for their decisions (at para 43). Beyond administrative decision-making, transparency is also given substance in access to information legislation at all levels of government. See: *Access to Information Act* RSC, 1985, c. A-1; *Freedom of Information and Protection of Privacy Act*, RSO 1990, c F 3.

[7] Daniel Wyatt, "The Many Dimensions of Transparency: A Literature Review" (2018) Helsinki Legal Studies Research Paper Series, No 53.

[8] This of course presumes that there is information that is capable of being understood. An assumption that may not hold up to much scrutiny in the case of automated decision-making. This critique is developed further in part 4.1.

relevant decision-making body."[9] Translated to the context of automated decision-making and profiling, we might think of decision rules, data, false positives and negatives, and features as the rationale and the actual construction of the automated system in terms of feature engineering, class specification, and algorithmic selection as the process. And while we might seek to understand either the rationale or the process behind an automated system as an end in and of itself, we may also seek that understanding because it allows us to hold a decision-maker accountable. That is, we expect that if provided access to the relevant information, we would be capable of understanding or challenging that decision. Boiling it down, transparency is a condition that allows individuals to access information that permits them to understand, and if necessary, hold accountable an organization responsible for using his or her personal information in an automated decision- making or profiling system.[10]

In principle, this is a laudable policy proposal. In practice though, there are three sets of problems with applying this logic of transparency to automated decision-making and profiling. Although not all totally unique to the automated nature of these processes, I aim to demonstrate how these three limits combine to undermine and in some cases hollow out an individual's ability to leverage transparency to understand or challenge how her personal information is used.

The first of these limits is in relation to the type of information that individuals would glean from an organization who was transparent about their automated decision-making and profiling system. Even if we define it as information that allows an individual to understand or challenge an

---

[9] Patrick Birkinshaw, 'Freedom of Information and Openness: Fundamental Human Rights?' (2006) 58 Administrative Law Review 177 at 189.

[10] As indicated by case law in *Dagg v Canada* (Minister of Finance), [1997] 2 SCR, dissenting, 403 at para 68, Canada (Information Commissioner) v Canada (Transportation Accident Investigation and Safety Board), 2006 FCA 157; *Canada (Information Commissioner) v Canada (Commissioner of the Royal Canadian Mounted Police),* [2003] 1 SCR 66, 2003 SCC 8, at para 23, the definition of Personal Information must be given a broad and expansive interpretation.

automated decision or instance of profiling, we are left asking: meaningful according to who? As Amitai Etzioni notes, the idea of transparency "assumes that those who receive the information released by producers or public officials can properly process it and that their conclusions will lead them to reasonable action."[11] The problem with this assumption though is that people differ in age, technical and legal literacy, time available to spend interacting with explanatory interfaces, and knowledge about avenues of institutional recourse (e.g. how to contest a decision). Lessons from privacy notices illustrate that when meaningful information is pegged to a prototypically reasonable person, the utility of such information is crippled in the eyes of vast swaths of the population. Considering the elevated levels of technical complexity involved in automated systems, I am especially concerned that a rigid standard of meaningful information will hamper our ability to understand or challenge them.

The second problem is that it's not clear that that transparency should always trump secrecy with respect to providing individuals with direct access to information about automated decision-making and profiling systems. Rather, there are times where secrecy-promoting concerns can justify an organization's refusal to open up the proverbial "black box" to anyone who requests access. These include an organization's desire to prevent gaming, to protect their confidential information, and the public's desire to not have their own personal information revealed in the process of granting someone else access.[12] Every time any number of these secrecy-promoting concerns exist, an individual's ability to access information and therefore understand or challenge an automated decision-making or profiling system is further hollowed out.

[11] Amitai Etzioni, 'Is Transparency the Best Disinfectant?' (2010) 18 The Journal of Political Philosophy 389 at 398.
[12] There is also potentially a concern about how providing individuals with access to information about an automated system's inferences may lead to stigmatization. However as part 4.4 describes, this worry cannot justify secrecy over transparency.

The third problem is that from a technical point of view, machine-learning techniques which frequently underlie these automated systems often generate outputs with underlying logics that are random, dynamic, non-intuitive, and multi-dimensional. If the information generated under the guise of "meaningful information about the logic involved" in an automated decision-making or profiling system is inscrutable on any or all of these grounds, transparency may not be capable of generating understanding or accountability. Moreover, turning up the dial on explainability may come at the cost of an automated system's performance.[13] If transparency results in, for example, sub-optimal lending decisions on account of having to reduce the model's complexity, we actually might not want to demand that automated decision-making and profiling systems permit individuals to understand or challenge them directly.

So, while I agree that Canada should move in the direction of regulating automated decision-making and profiling, these three limits indicate that a regulatory regime rooted in a desire for increased transparency is by itself insufficient to achieve its stated aims. In my view, there are two revisions that can be made to overcome these limits. The first is that we must move beyond a one size fits all approach to the definition of meaningful information in order for transparency to be capable of yielding understanding or accountability. This means defining "meaningful information" as that which permits an individual to understand or challenge an automated decision or instance of profiling having due regard for variability in age, technical and legal literacy, time, and knowledge about avenues of institutional recourse. This doesn't mean that information needs to be tailored to each specific individual on a case by case basis. Rather, it means that information must be provided at a number of levels of complexity and leverage insights from disciplines like human-computer interaction and user-experience design to employ audio-visual tools and data

[13] Doshi-Velez et al, "Accountability of AI Under the Law", *supra* note 4.

visualization techniques to facilitate its comprehension. The goal here is to deliver information that is subjectively and instrumentally meaningful to the widest possible audience in the sense that it permits them to understand it and use it to generate accountability. However even with this revision in place, there are going to be cases where secrecy-promoting concerns and the characteristics of machine-learning make it either impossible or undesirable to provide individuals with direct access to meaningful information about an automated system. Notice though, that these are really only limits when we try to deliver information directly to individuals. If instead we were to shift the recipient of transparency, as in access to meaningful information, over to a neutral third party, individuals might be robbed of the opportunity to understand these automated systems, but they could take solace in knowing that those systems are being verified and held accountable. Therefore, my second recommendation is that when organizations are unable to provide individuals with direct access to information on account of secrecy-promoting concerns or the characteristics of machine-learning systems, they must have their automated decision-making and profiling systems pre-emptively verified by a certification body housed within the Office of the Privacy Commissioner (OPC). With both of these revisions in place, individuals will be in a better place to understand and challenge these automated systems themselves or when this is not possible, trust that these systems are still being held accountable.

To get here, this article is separated into six parts. In part two, I demonstrate how in principle, there are two strong justifications for wanting transparency in the form of an individual access right with respect to automated decision-making and profiling. First, transparency is useful for allowing individuals to understand how their personal information is being used and how it affects their interactions with these automated systems. Such an understanding is important in the post-enlightenment sense, as it is necessary for illuminating how automated systems might restrict

autonomy and self-determination. This is especially relevant in a world where automated systems are increasingly involved in shaping what we know and how we understand ourselves and our social relationships. Second, transparency can be instrumentally useful insofar as it reveals information that enables an individual to challenge the organization responsible for the automated system. In this sense, transparency is also a mechanism for generating accountability. To illustrate how, I explore how transparency can be buttressed to reveal if an automated decision-making system violated privacy law's purpose specification principle or if it produced a discriminatory output.

In part three, I explore how the GDPR goes about operationalizing transparency with respect to automated decision-making and profiling. In particular, I establish the scope of automated decision-making and profiling activities the GDPR is concerned with and explain how it conceives of meaningful information. This section is primarily descriptive and serves the function of setting up my own recommendations for regulating automated decision-making and profiling in Canada discussed in the following parts.

In part four, I explain that even though there are solid theoretical justifications for wanting transparency, in practice, it is not clear that we can always rely on it to realize any of these principled ambitions. In other words, transparency as described in part two and in the GDPR, will not always be capable of permitting an individual to understand or challenge an automated decision-making or profiling system. To make this argument, I discuss in more detail how we can expect the three limits of transparency – epistemic variability, secrecy-promoting concerns, and the characteristics of machine-learning – to commingle and undermine an individual's ability to simply look under the hood of an automated decision-making system to understand it or to hold its progenitors accountable.

Having identified these limits, part five advances two recommendations for how a Canadian approach to regulating automated decision-making and profiling in the private sector ought to overcome them. As already identified, the first of these concerns landing on a definition of meaningful information that requires organizations to take into account epistemic complexity. The second is that where an organization's use of an automated decision-making or profiling system makes them reluctant or unable to provide meaningful information to individuals directly, that organization must have their system pre-emptively verified by a certification body housed in the OPC before it can be used. I develop this recommendation by pointing to a number of technical and organizational methods that can form the basis of this certification process.

I bookend my discussion in part six by flagging and responding to two likely objections to these recommendations and by identifying one open question. First, I consider whether requiring organizations to cater to a wide variety of epistemic needs would be unduly costly and whether those costs would chill innovation disproportionately among small to medium size enterprises. Second, I consider whether the OPC is the appropriate venue for a certification body or whether the private sector would be better equipped to fill this role. Finally, I flag that whether a certification body exists within the OPC or at an arms-length distance in the private sector, we may find ourselves back in a position of demanding, but on account of the same limits discussed here unable to get, transparency over the decisions made by this certification body.

I conclude by bringing together the paper's recommendations into a coherent whole, submitting that if *PIPEDA* is amended to regulate solely automated decision-making and profiling which produces legal or similarly significant effects, the limits of this transparency-based regulatory approach must be accounted for.

Part 2: Two Justifications for Transparency: Understanding and Accountability

It's first worth considering why would we want organizations to be transparent about how our personal information is used in the process of automated decision-making and profiling in the first place. In my view, there are two main justifications for wanting transparency over these systems. First, access to the inner workings of that system where the processing of personal information takes place would, at least in principle, permit an individual to better understand their interactions with that automated system. Such an understanding is instrumentally useful insofar as it exposes how our experiences are mediated by decision-making and profiling systems. On a deeper level, such an understanding allows us to identify and mitigate against the forces that bear on our epistemological and ontological development and is therefore important to bolstering autonomy and self-determination. Second, if, based upon this understanding, a person suspects that the decision-making or profiling system has run afoul of a legal rule it would permit him or her to challenge and hold the organization responsible for it accountable. In other words, the conditions brought about by a state of transparency are important insofar as they enable an individual to challenge the decision or instance of profiling. Again, on a deeper level, where transparency generates accountability it reinforces intrinsically valuable ends like privacy and equality. In the discussion that follows, I unpack in more detail how generating transparency, in the form of an individual access right, over automated decision-making and profiling would allow an individual to understand or challenge an automated decision or instance of profiling. To be clear, understanding and accountability are what transparency should deliver on in principle – the limits of this thinking are developed in part four.

2.1 : Transparency and Understanding: Autonomy and Self-Determination

The reason for wanting to understand an automated decision-making or profiling system is not only to challenge it if it appears to have violated a legal rule. Rather, by itself, understanding an automated system is also useful for identifying and mitigating against the forces that bear on our epistemological and ontological development.[14] This is vital in in an era of near ubiquitous computing, where automated decision-making and profiling systems are continually mediating our experience and affecting how and what we know about the world and ourselves. Consider the significance the first ten results or autocomplete suggestions delivered by a Google search has on what you are likely to (not) know about most things and how those results inform your knowledge of yourself as well as your knowledge of others. Without a meaningful understanding of the mechanisms that govern the construction of these realities, "the formation of our will power is steered by what we cannot assess."[15] Therefore, insofar as our epistemological and ontological development is affected by automated decision-making and profiling systems, autonomy and self-determination are implicated.[16] Transparency would be an antidote to this reduction in autonomy and self-determination in the sense that it would expose to individuals an array of forces that bear on their epistemic and ontological development. Instrumentally, such an understanding would allow an individual to calibrate their reactions to the outputs of these systems and potentially temper their influence. According to this line of thought, understanding the connection between

[14] This argument should not be foreign to privacy scholars, whose discipline is often justified on the basis of Locke's theory of autonomy and self-ownership. Indeed, the EU's move to inject transparency into automated decision-making systems through its data protection laws indicates that transparency is cut from the same normative cloth as privacy. The proximity of this justification to the predominant theoretical foundation of privacy provides partial support for regulating automated decision-making within the ambit of Canada's privacy framework. This is not the *only* good theoretical justification for privacy though, see: Ari Ezra Waldman, "Privacy as Trust: Sharing Personal Information in a Networked World" (2015) 69:3 University of Miami Law Review 559.

[15] Mireille Hildebrandt, "The Dawn of a Critical Transparency Right for the Profiling Era" in Jacques Bus et al, eds *Digital Enlightenment Yearbook 2012* (Amsterdam: IOS Press, 2012) 41 at 47.

[16] See: Natascha Just & Michael Latzer, "Governance by algorithms: reality construction by algorithmic selection on the Internet" (2017) 39:2 Media, Culture, and Society 238; John Danaher, "The Threat of Algocracy: Reality, Resistance and Accommodation" (2016) 29:3 Philosophy and Technology 245; Sang Ah Kim, "Social Media Algorithms: Why You See What You See" (2017) 2 Geo Law Tech 147.

our personal information and an automated system would bolster individual autonomy and self-determination defined as the relative absence of constraints on one's epistemological and ontological development.

2.2 : Transparency and Accountability: Privacy and Discrimination

Sometimes in the course of an individual coming to understand an automated decision-making or profiling system's use of her personal information, she will discover that it is likely in violation of a legal rule. Therefore, insofar as looking 'under the hood' of an automated system reveals a set of facts that indicate the legal efficacy of that system, transparency is a gateway to generating accountability.[17] The following discussion explores the relationship between transparency and accountability as it might pertain to privacy and discrimination law. In the former, a state of transparency allows an individual to challenge an automated decision or instance of profiling on the grounds that it violated the purpose specification principle by drawing unreasonable inferences. In the latter, it allows an individual to root out discrimination within an automated decision-making or profiling system by learning for example, that an automated decision relied on a proxy, or redundant encoding of, a prohibited ground of discrimination. I discuss each in turn.

2.2.1 : Transparency, Accountability, and Privacy

[17] Whether transparency actually leads to accountability is an empirical question on which the jury is still out. For a review of the literature on the theoretical and empirical relationship between transparency and accountability, see: Wyatt, "The Many Dimensions of Transparency: A Literature Review", *supra* note 7.

Increasingly, data mining and machine-learning techniques permit decisions to be made on the basis of inferences that completely confound human comprehension.[18] For instance, by knowing discrete and apparently unconnected facts about you – your shirt colour, gait, driving habits, and the e-mail font you use – companies could, using algorithms that sort the profiles of hundreds of thousands of people like you, accurately predict whether you are a good credit risk.[19] Our mutual confusion on whether this is possible or not is the point. As Solon Barocas notes, data mining and machine-learning "radically redefines what it means for something to be relevant."[20] Because the inferences that can be drawn from data are often impossible to predict, automated decision-making and profiling problematizes privacy law's aspiration to make clear the purposes for which personal information is used up front. As a fair information principle, purpose limitation is considered necessary to legitimize an individual's consent at the outset of data processing.

Not all privacy scholars have given up on the apparent incompatibility between purpose limitation and data mining though. Some suggest that we can draw lines around the types of inferences it is reasonable for data mining applications to make on the basis of an individual's consent, leaving the inferences that fall outside these lines in violation of the purpose specification principle. For example, Peter Leonard suggests that surprising or disturbing inferences should be

[18] Quoting Usama Fayyad, data mining is the "mechanized process of identifying and discovering useful patterns, models, and relations in data." Usama Fayyad, "The Digital Physics of Data Mining" (2001) 44:3 COMM ACM 62. Machine-learning takes this one step beyond the descriptive function and seeks to to build models of what is happening behind some data so that it can predict future outcomes. When these predictions are acted upon, then an automated decision has been made. Already, automated decision-making is committed to a complicated relationship with truth. That is, it relies on (1) historical, group level (2) patterns, and relations in data that (3) need not be causal but must only correlate to a sufficient degree (4) in order to be *useful* according to the amoral purpose it's put. Whether one sees a connection between automated decision-making and truth at all lies in part in their stance towards David Hume's problem of induction.

[19] Jeffery Rosen, "Who Do Online Advertisers Think You Are?" *New York Times* (November 30, 2012) online: https://www.nytimes.com/2012/12/02/magazine/who-do-online-advertisers-think-you-are.html; Kathy, O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York City, NY: Penguin Random House, 2016).

[20] Solon Barocas, "Panic Inducing: Data Mining, Fairness, and Privacy" (2015) PHD Thesis, New York University at 110.

restricted.[21] However, not all inferences that are surprising are problematic. Rather, we may be pleasantly surprised by what data mining reveals to us. Alternatively, Jason Millar has suggested that inferences are a problem when they implicate what he calls 'core privacy', defined as an observation of "an individual's unexpressed psychological properties to which only the individual has first-person access, and that are not knowable by anyone else, except by the individual's prior divulgence of them, or by an unreasonable inference based on other facts already known about the individual."[22] However, the problem with this approach is that data mining and machine-learning also permit inferences that are unreasonable.[23] In fact, these methods are often useful precisely because they permit the discovery of patterns that confound our expectation of reasonable inference.

Like Millar and Leonard, Barocas agrees that the problem with inference is that it provides an indirect route to access sensitive information.[24] However, building on Helen Nissenbaum's theory of contextual integrity, he slightly departs from these authors, suggesting that inferential discoveries are most problematic when they subvert the social and contextual norms that regulate how and when inferences can be made.[25] For example, inferring someone's political belief on the basis of a magazine purchase comports with our social and cultural expectations of what a magazine purchase might reveal about ourselves. However, there are less well defined social or cultural expectations surrounding what a basket of groceries would say about our political beliefs.[26] According to this formulation, inferences are most problematic not when they are

[21] Peter Leonard, "Customer Data Analytics: Privacy Settings for 'Big Data' Business" (2014) 4:1 International Data Privacy Law.
[22] Jason Millar, "Core Privacy: A Problem for Predictive Data Mining" in Ian Ker et al, *Lessons from the Identity Trail:* Ano*nymity, Privacy and Identity in a Networked Society* (Oxford: Oxford University Press, 2009) 103.
[23] Barocas, "Panic Inducing", *supra* note 20 at 119.
[24] Barocas, "Panic Inducing", *supra* note 20 at 113.
[25] Barocas, "Panic Inducing", *supra* note 20 at 136.
[26] Thought experiment borrowed from Barocas, "Panic Inducing", *supra* note 20 at 129.

surprising or unreasonable but when they are divorced from logical, cultural, and social contexts from which the information was first collected.

Let me now connect this to my argument that transparent automated decision-making or profiling system can be used by individuals to generate accountability in privacy law. Suppose an individual was granted access to a range of inferences that permitted a decision or profile to be made about her. Through access to these inferences, she would be in the position to contrast them with the purposes stated at the beginning of the data processing cycle. To the extent that these inferences are made on the basis of information that people have never understood as a social or cultural signal of such a quality, she could have reason to challenge the decision on the grounds that her information was used in a way that is not consistent with the purposes for which she originally consented.

2.2.2 : Transparency, Accountability, and Discrimination

The second area where we might expect transparency to generate accountability is with respect to discrimination law. This might be the case if individuals were granted access to inspect the data as well the rules the automated decision-making or profiling system learned inductively through those data. First, access to the data used to train a supervised machine-learning algorithm may reveal a sample bias in those data. For example, if an automated resume checking system searching for a new CEO was trained on the resumes of previous CEOs, the resulting model might be skewed by a sample dominated by men and could mistakenly rely on gender as a proxy for "Good CEO". The past is not always a good indicator of what the future ought to be and we can't let our normative commitments such as that to equality become derailed by a new technological way of operating. If we do, and we permit these highly consequential decisions to be made in the

dark, we may be doomed to replicate a "self-reinforcing and self-perpetuating system, where individuals are forever burdened by a history that they are encouraged to repeat and from which they are unable to escape."[27] Transparency here, in the form of access to the training data of an automated decision-making system might reveal the sample bias in those data, making it instrumentally useful insofar as it can be acted upon to facilitate a legal challenge on those grounds. Second, access to the decision rules learned inductively through those data could reveal that an automated system relied upon a rule which was a proxy for, or 'redundant encoding' of, a prohibited ground of discrimination.[28] For instance, rather than relying explicitly upon somebody's race to deny them a credit extension, an automated system could discover that people who share the same postal code as that individual have been financially risky applicants in the past. But we also know that someone's postal code can be correlated with race.[29] If that individual were to discover that he was equal to other applicants who applied for the same credit extension in all other aspects except for his postal code, a case could be made that the automated decision-making system was discriminatory. Transparency here, in the form of access to decision rules may permit individuals to hold the organizations who use these systems accountable to the law and buttress our commitment to equality.

Concluding Remarks on the Justifications for Transparency

Fundamentally, a transparent automated decision-making or profiling system is one whose rationale and processes are open to view. There are two instrumental justifications for wanting

[27] Solon Barocas et al, "Governing Algorithms: A Provocation Piece" (2013) Paper prepared for the "Governing Algorithms" conference at New York University.
[28] For example, see: Julia Angwin et al, "Machine Bias" *Pro Publica* (23 May 2016) online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[29] Michael Veale & Reuben Binns, "Fairer Machine-learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data" (2017) 4:2 Big Data & Society 1 at 4.

transparency of this sort, both of which are only one step removed from a deeper intrinsic purpose. First, transparency serves an instrumental pedagogical function by permitting individuals to understand how an automated decision or instance of profiling mediates their epistemological and ontological development. In other words, a state of transparency enables self-determining and autonomous agents to know what external forces colour how and what they know about the world, themselves, and others.[30] Second, transparency is instrumentally useful insofar as it permits an individual to access information which can be used to generate accountability. For instance, a state of transparency could allow an individual to challenge an organization on the basis that its automated system relied on inferences that violated the purpose specification principle because they were made in the absence of socially meaningful cues and therefore could not have been consented to. Also, it could allow an individual to root out discrimination within an automated decision-making or profiling system by allowing him to find out that an automated decision relied on a redundant encoding of a prohibited ground of discrimination. In this sense, transparency is important for advancing equality. As the next section goes on to describe, there is significant consilience between this understanding of transparency and the EU's approach to regulating automated decision-making and profiling in the GDPR.

Part 3: Regulatory Frameworks to Operationalize Transparency: GDPR

The European Union's General Data Protection Regulation (GDPR) is by far the most commonly discussed regulatory framework in relation to bringing about the transparency of automated decision-making and profiling.[31] It is relevant to this discussion for two reasons. First,

[30] See: Tarleton, Gillespie, "The Relevance of Algorithms" in Tarleton Gillespie et al, eds, *Media Technologies Essays on Communication, Materiality, and Society* (Massachusetts: MIT Press, 2014) 167.
[31] Tal Zarsky, "Incompatible: The GDPR in the Age of Big Data" (2017) 47 Seton Hall Law Review 995; Bryan Casey et al, "Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of

it allows me to expound upon the language used by the GDPR to establish the scope of automated decision-making and profiling activities subject to regulation. In particular, I describe what it means for a decision to be "solely" automated, I describe how the GDPR establishes a qualified prohibition on solely automated decision-making as well as solely automated profiling, and what it means for a decision to produce legal or similarly significant effects. Later in the paper when offering my recommendations as to how Canada should regulate in this space, I refer back to this regulatory language, arguing that the GDPR is a good model for Canada to follow when establishing the scope of automated decision making and profiling activities that are subject to regulation. The second reason the GDPR is relevant to my discussion is because it tracks relatively closely to part two's discussion of the theoretical justifications for transparency. That is, the GPDR is a regulatory framework which intends to create the conditions that allow individuals to access information that would permit him or her to understand and challenge an automated decision or instance of profiling. This can be seen in how it conceives of the role of providing individuals "meaningful information" about the logic involved in automated decision-making and profiling activities. The purpose of extracting this particular interpretation is to set up my discussion in part four where I highlight three classes of limits to transparency so understood.

3.1 : What Types of Automated Activities Generate Transparency Requirements?

3.1.1: A Decision Based "Solely" on Automated Processing

Algorithmic Audits in Enterprise" (2018) Forthcoming in Berkley Technology Law Journal; Sandra Wachter et al, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2016) 7:2 International Data Privacy Law 76; Maja Brkan, "Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond" (2018) submitted for 'Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence' A previous version of this paper was published as Brkan, M. (2017), 'AI-supported decision-making under the General Data Protection Regulation', in Proceedings of the 16th International Conference on Artificial Intelligence and Law, London; Merle, Temme, "Algorithms and Transparency in View of the New General Data Protection Regulation" (2017) 4 European Data Protection Law Review 473.

Pursuant to a joint reading of Articles 22(1), 13(2)(f), 14(2)(g), and 15(1)(b), the GDPR only requires organizations to generate meaningful information about the logic involved in "solely" automated decisions.[32] However, the GDPR does not offer a definition of the word solely which is somewhat problematic given that many systems today involve a combination of human and machine decision-making where human involvement falls along a spectrum rather than a bipartite axis. In an article titled "Mapping the Logical Space of Algocracy", John Danaher describes it thusly:[33]

	(1) Humans perform task	(2) Task is shared with algorithm	(3) Algorithms perform task; Humans supervise	(4) Algorithms perform task; No human input
Sensing	Y or N?	Y or N?	Y or N?	Y or N?
Processing	Y or N?	Y or N?	Y or N?	Y or N?
Acting	Y or N?	Y or N?	Y or N?	Y or N?
Learning	Y or N?	Y or N?	Y or N?	Y or N?

According to Danaher's formulation there are 256 possible combinations of human and algorithmic cooperation.[34] Should data controllers be permitted to dodge the regulation's application and the transparency guarantees that flow from it by pointing to any modicum of human involvement at any stage, the larger project of generating transparency would be routinely averted. Fortunately, guidance from the Article 29 Data Protection Working Party Guidelines on

[32] GDPR, *supra* note 5 at art 22(1), 13(2)(f), 14(2)(g), 15(1)(h).
[33] John Danaher, "Mapping the Logical Space of Algocracy", (2015) *Philosophical Disquisitions*, online: http://philosophicaldisquisitions.blogspot.com/2015/06/how-might-algorithms-rule-our-lives.html.
[34] We can represent an analog procedure – one in which all the decision-making tasks are performed by humans – like this: [1, 1, 1, 1].

Automated Individual Decision-Making and Profiling (Article 29 WP) and the UK's Information Commissioner's Office (ICO) helps disambiguate what it means for a decision or instance of profiling to be solely automated. Using Danaher's illustration, the GDPR appears to be concerned primarily with what happens in the 'acting' stage – where a decision is made, or a profile created. This conclusion is drawn from a reading of the Article 29 WP guidelines on automated decision making and profiling which despite not constituting a binding interpretation of the GDPR indicates that a decision will be considered based solely on automated processing unless a human is able to exert real influence on the outcome of the decision.[35] Moreover, that intervention must be more than a "token gesture" and must be made by someone who is technically and legally competent and has the authority to actually affect the outcome of the decision.[36] This position is in congress with a separate opinion released by the UK ICO where it opined that even where nominal human intervention formally 'takes' the decision, Article 22(1) should still apply.[37]

In addition to these legal opinions, social science research into automation bias gives us good reason to be suspect of how much of an influence a 'human in the loop' is actually capable of exerting over an otherwise automated decision.[38] Automation bias refers an instinctive proclivity to cede control to and accept the conclusions made by automated decision support systems,[39] and our "tendency to disregard or not search for contradictory information in light of a computer-generated solution."[40] Researchers have found that it occurs in both naïve and expert

[35] Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679, at 10 [Article 29 WP].
[36] Article 29 WP, *supra* note 35 at 10.
[37] Information Commissioner's Office, Feedback request – profiling and automated decision-making, (2017) online: https://ico.org.uk/media/2013894/ico-feedback-request-profiling-and-automated-decision-making.pdf at 19.
[38] Meg Leta Jones, "The right to a human in the loop: Political construction of computer automation and personhood" (2017) 47:2 Social Studies of Science 216.
[39] Raja Parasuraman et al, "Humans and Automation: Use, Misuse, Disuse, Abuse" (1997) 39:2 Human Factors 230.
[40] Mary Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems" (2004) AIAA 1st Intelligent Systems Technical Conference at 1.

participants and cannot be prevented by training or instructions.[41] One thing that separates this phenomenon from regular confirmation bias is that our susceptibility towards automation bias can be exacerbated by the complexity of the decision support system.[42] This particular feature of automation bias is especially relevant given the complexity inherent in many machine-learning techniques. Where these techniques are used, an individual's intervention in an automated decision is technically preempted by the system's non-intuitiveness and multi-dimensionality may have to be taken with a grain of salt. Altogether, there are just too many signals that indicate that the mere involvement of a human in an automated decision does not guarantee he or she exercises any meaningful influence on that decision. In accordance with this social science evidence and legal opinion, a decision or instance of profiling ought to be considered solely automated unless a human is capable of intervening in that process in a non-trivial way.

3.1.2 : A Right or Prohibition?

When an automated decision-making or profiling system is solely automated, its use can only be justified on the basis of the three exceptions: where it is necessary for entering into, or performance of, a contract between the data subject and a data controller; is authorized by a Union or Member State law to which the controller is subject which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or is based on the data subject's explicit consent.[43] In other words, the GDPR creates a qualified prohibition on "fully automated individual decision-making, including profiling that has a legal or similarly significant

[41] Raja Parasuraman & DH Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration" (2010) 52:3 *Human Factors* 381 at 381.
[42] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems", *supra* note 40 at 1.
[43] GDPR, *supra* note 5 art 22(2)(a), (b), (c).

effect."[44] When the automated decision-making or profiling is justified on any of these grounds, the requirement to provide individuals with access to meaningful information about the logic supporting the automated activities kicks in. The last thing worth noting here, is that the GDPR also enacts a qualified prohibition on solely automated profiling even where a decision has not been made. This interpretation stands in contrast to that of Isak Mendoza et al., who expressed an opinion that as a matter of law Article 22 enacts a qualified prohibition on decision-making made on the basis of profiling but not profiling in the absence of a decision.[45] However, the Article 29 WP has issued guidance suggesting that "the GDPR does not just focus on the decisions made as a result of automated processing or profiling. It applies to the collection of data for the creation of profiles, as well as the application of those profiles to individuals."[46] On this basis, we can be confident that the GDPR enacts a qualified prohibition on solely automated decision-making and solely automated profiling whether or not that profiling was used to inform a decision.

3.1.3 : Legal or Similarly Significant Effects

Further constraining the GDPR's application is a requirement that the decision or profiling produces "legal effects concerning him or her or similarly significantly affects him or her."[47] While the realm of legal effects can be uncontroversially described as those where legal status is altered or legal duties are created, when a decision 'similarly significant affects' someone is less straightforward. Recital 71 of the GDPR tries to describe these cases as decisions that affect one's

[44] Article 29 WP, *supra* note 35 at 12.
[45] Isak Mendoza & Lee Bygrave, "The Right Not to be Subject to Automated Decisions Based on Profiling" in Eleni Synodinou et al, eds EU Internet Law: Regulation and Enforcement (Switzerland: Springer International Publishing, 2017) 77.
[46] Article 29 WP, *supra* note 35 at 6; Profiling here, is defined as (1) any form of automated processing of personal data which (2) uses those data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements. GDPR, *supra* note 5 art 4(11).
[47] GDPR, *supra* note 5 art 22.

personal preferences, interests, reliability or behaviour.[48] However nobody exists, deliberates, or acts in a vacuum. Instead we are constantly awash in external forces, which to varying degrees shape our preferences, interests, reliability, and behaviour. When this mediation turns from being benign to being corrosive of 'free will' or constrains an individual's choice such that it has a significant effect on his or her life and calls for legal intervention is far from clear.[49] To further complicate the equation, different people are liable to respond to the mediation of their preferences and interests in different ways and it is not clear whose subjectivity our standard of 'significantly affects' should appeal to. Finally, the significance of any instance of automated decision-making or profiling may reveal itself over time and in conjunction with many instances.[50]

Given the real potential for almost anything to matter, we might want to subject a wide number of practices – advertising, content recommendation,[51] price discrimination, etc. – to regulatory intervention.[52] However, the further down this rabbit hole of algorithmic influence we venture, the more at risk we are of overregulating a massive and largely innocuous subset of socio-technical life. Insofar as there are costs imposed by this regulatory regime, widening its scope is unadvisable. To note just one example, the more automated activities we capture in this language, the more organizations we risk burdening with the upfront technical and organizational costs involved with 'doing' transparency.[53] Besides the financial costs conferred by a malleable

[48] GDPR, *supra* note 5 recital 71.
[49] To take another position, if research at the intersection of philosophy of mind and neuroscience which casts doubt on the extent to which we have libertarian or contra-causal free will is right,[49] and we are just a sum of our influences, then there is really no end to the catalogue of automated decision-making and profiling that we might want to interrogate. See: Sam Harris, *Free Will* (New York: Free Press, 2012).
[50] Jatinder Singh et al, "Decision Provenance Capturing data flow for accountable systems" (2018) Computers and Society online: arXiv:1804.05741v1.
[51] Lucas Introna & Helen Nissenbaum "Shaping the Web: Why the Politics of Search Engines Matter" (2000) 16 The Information Society 169.
[52] Especially considering the slow but steady arrival of 'smart cities'. See Matthew Jewell, "Contesting the decision: living in (and living with) the smart city" (2018) International Review of Law, Computers, and Technology.
[53] Matthew Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29:2 Harvard Journal of Law and Technology 354.

interpretation, such an approach may have the adverse effect of overwhelming individuals with information about every automated decision-making or profiling system they interact with on a daily basis. A similar tension exists with respect to data breach notifications. Currently, Canada's data breach notification regime is activated only when the information in question would pose a real risk of significant harm.[54] If the threshold that triggers a data breach notification is on the low end of 'real risk of significant harm', notifications may become more common and individuals may begin to ignore them even when they fall on the more serious end of the risk continuum. To avoid such an outcome, pegging 'similarly significant affects' to a reasonably high level of impact would instead reinforce to the public that when they are delivered information by an organization they can be sure that it's worth paying attention to.

In my view, the GDPR strikes a reasonable balance here. As a general rule, the more adverse the effects are and the closer they are in proximity to resembling legal significance, the more likely they will be considered similarly significant. This interpretation is also supported by a textual update from the preceding Data Protection Directive (DPD).[55] Article 15 of the DPD reads "Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him."[56] The GDPR however, adds the word "similarly" before significantly affects. We might read into this an intention on behalf of the drafters to draw a link between legal effects, and similarly, as in non-trivial, effects.[57] As indicated

[54] *Digital Privacy Act* SC 2015, c 32, (f); *PIPEDA,* s 4.3.4, *supra* note 1.
[55] Article 29 WP, *supra* note 35 at 10.
[56] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data [1995] OJ L 281/31 art 15 [DPD].
[57] Article 29 WP, *supra* note 35 at 10; Mendoza et al, "The Right Not to be Subject to Automated Decisions Based on Profiling", *supra* note 45 at 89.

by Recital 71, e-recruiting practices or automatic refusals of online credit applications are examples of automated decisions meeting this regulatory threshold.[58]

3.2 : Meaningful Information

With all of this in place – automated decision-making *or* profiling exists that is *solely* automated and produces *legal effects* or *similarly significant affects* him or her – an individual has the right to obtain from the data controller "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing."[59] How are we to interpret 'meaningful information'? According to a conference paper presented in 2016 by Bryce Goodman and Seth Flaxman, the GDPR provides a 'right to an explanation' of automated decision-making.[60] However, that idea was criticized most popularly by Watcher et al.[61] but also by Edwards & Veale among others.[62] Since then, some clarity has been restored to the likely scope of meaningful information by the likes of Andrew Selbst and Julia Powels,[63] as well as Gianclaudio Malgieri and Giovanni Comandé.[64]

Notwithstanding the important contributions of these authors, looking strictly at the GDPR and the Article 29 WP guidance provides some insight into the interpretation of meaningful information. While not a legally binding text, Recital 71 of the GDPR suggests that individuals

---

[58] GDPR, *supra* note 5 recital 71.
[59] GDPR, *supra* note 5 art 15
[60] Bryce Goodman & Seth Flaxman "A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection" (2016) ML and the Law (NIPS Symposium 2016) online: http://www.mlandthelaw.org/papers/goodman1.pdf.
[61] Wachter et al, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *supra* note 31.
[62] Lilian Edwards & Michael Veale "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18.
[63] Andrew Selbst & Julia Powels, "Meaningful Information and the Right to Explanation" (2017) 7(4) International Data Privacy Law 233.
[64] Gianclaudio Malgieri & Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation" (2017) 7:4 International Data Privacy Law 243.

ought to be provided with information that that is sufficiently comprehensive to allow them to understand and challenge the decision.[65] The Article 29 WP backs this up, opining that "the data subject will only be able to challenge a decision or express their view if they fully understand how [a decision] has been made and on what basis."[66] This may mean, in the context of explaining the rejection of a credit application, that the decision-maker should provide the rationale behind, the criteria relied on in reaching the decision, and the source of those data.[67] This could include an explanation of the relevance of a credit score to making fair and responsible lending decisions, the main characteristics considered in reaching the decision, as well as the source of that information and its relevance.[68] This conception tracks well with part two's discussion of the instrumental importance of transparency.[69] That is, in order for information to be meaningful, an individual must be able to understand and act upon it. Whether that be decision rules, counterfactual explanations,[70] anonymized data, algorithms, or a rationale for automating that decision, all information can be measured against this instrumental function.[71] What isn't clear though, is the GDPR's stance on the question, meaningful information according to who? Is meaningful information a one-size fits all, privacy policy-esque document meant to cater to the middle of the reasonable person, or is it more flexible? As part four goes on to describe, I believe that whether

[65] GDPR, *supra* note 5 Recital 71. Like the Article 29 Working Party Guidance, recitals are not binding but are meant to function as interpretative aids especially when ambiguity exists in the main text of the regulation. See: *Case 215/88 Casa Fleischhandels* [1989] European Court of Justice ECR 2789 [31]; Roberto Baratta, 'Complexity of EU Law in the Domestic Implementing Process' (2014) 2 The Theory and Practice of Legislation 293 at 17.
[66] Article 29 WP, *supra* note 35 at 16.
[67] Article 29 WP, *supra* note 35 at 14.
[68] Article 29 WP, *supra* note 35 at 26.
[69] It also tracks with how Selbst & Powels conceive of the term: "Meaningful Information and the Right to Explanation", *supra* note 63 at 7.
[70] Sandra Wachter et al, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" (2018) Forthcoming in the Harvard Journal of Law and Technology; Tim Miller "Explanation in Artificial Intelligence: Insights from the Social Sciences" (2017) online: https://arxiv.org/abs/1706.07269.
[71] Selbst & Powles, "Meaningful Information and the Right to Explanation", *supra* note 63 at 7.

transparency enables an individual to understand or challenge a decision or profiling largely turns on the answer to this question.

Concluding Remarks on how the GDPR Delivers Transparency

To summarize, the GDPR generates a *qualified prohibition* on *solely* automated decision-making *including profiling* which produces *legal or similarly significant* effects. When these automated activities are justified, for example on the basis of an individual's consent, he or she is entitled to request *meaningful information* about the logic involved in that automated decision or profiling activity. In principle, it would appear that enabling an individual to access to meaningful information should deliver on the instrumental value of transparency. That is, the GDPR should allow an individual to understand and potentially challenge that automated decision or instance of profiling. However, as the next section goes on to detail, there are three classes of limits which call into question whether transparency and its operationalization in the GDPR is capable of delivering on either of these objectives.

Part 4: Limits to Transparency

The last section described how the GDPR can be thought of as a transparency framework that provides individuals access to meaningful information about the logic involved in a solely automated decision or profiling system which produces legal or similarly significant effects. Because meaningful information is likely to be defined as that which permits an individual to challenge a decision or instance of profiling, the GDPR appears to track well with my discussion of the justifications for transparency in part two. In principle, this would appear to be a good thing. However, as this section demonstrates, there are three classes of limits which suggest that

operationalizing transparency in this way is not likely to empower an individual to do much understanding or challenging at all.

First, transparency presupposes the existence of an idyllically rational, motivated, technically literate, and epistemically homogenous audience to understand and act upon information about the logic involved in an automated decision-making or profiling system. Second, the types of information that individuals would have to gain access to in order to understand or challenge an automated decision or instance of profiling may permit them to game these systems, may reveal the organization's confidential information, may run the risk of revealing the personal information of others, and could raise concerns around stigmatization.[72] While the first two classes of limits may plague the efficacy of transparency in both analog and automated processes, the following limit is mostly unique to automated decision-making and profiling. That is, transparency is undermined in a technical sense by algorithmic systems that utilize machine-learning techniques which can generate random, dynamic, non-intuitive, and multi-dimensional decisions that can be difficult if not impossible for humans to parse, much less understand or challenge.[73] Considering the amount of interest in using AI-based techniques in all manners of profiling and decision-making activities, this limit in particular is likely to be the most difficult to overcome. When any or all of these limits are present, it's not clear that transparent automated systems will be of much use to individuals at all. That is, any explanation will be so hollowed out that it will typically be an unreliable vehicle for enabling an individual to understand or challenge the basis of an automated decision or instance of profiling. Ultimately, it is only by

[72] For example, see: Latanya Sweeney, "Discrimination in Online Ad Delivery" (2013) 11:3 ACM Queue 1.
[73] For example, see: Jon Christian, "Why Is Google Translate Spitting Out Sinister Religious Prophecies?" *Motherboard Vice* (July 20, 2018) online: https://motherboard.vice.com/en_us/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies.

engaging with the limits of transparency that we can chart a coherent and effective regulatory solution to this multi-dimensional problem in part five.

4.1 : Meaningful Information and Epistemic Complexity

We might find that efforts to generate transparency over automated decision-making in the form of a universal explanation might fail on account of individuals having different ideas of what constitutes meaningful information as well as their constraints on time, motivation, technical and legal literacy, and knowledge of avenues of institutional recourse. An analogy to a well-trod critique of privacy law will help me demonstrate why I suspect this will be the case. Scholars like Daniel Solove and Bruce Schnier have criticized whether privacy notices are capable of effectively informing individuals about how their personal information is likely to be collected, used, and disclosed.[74] Here, organizations distribute a generic document that individuals are supposed to understand and use to negotiate their privacy relationship with that organization. In reality though, privacy notices can not only be difficult for a 'reasonable person' to digest but individuals are often unmotivated to engage with them because they either lack the time or feel as though they lack the bargaining power to negotiate the terms of that agreement. In this context, it is not clear that being open and transparent with respect to the collection, use, and disclosure of personal information is that instrumentally useful to individuals at all. Nonetheless, privacy law still rests on an assumption that by providing this notice, individuals can successfully manage their own privacy relationships by either granting or withholding consent.[75]

---

[74] Oft cited in this context is that according to one study conducted in 2008 estimated, it would cost $781 billion in lost productivity if everyone were to read every privacy policy at websites they visited in a one-year period see: Aleecia M McDonald & Lorrie Faith Cranor, "The Cost of Reading Privacy Policies", (2008) 4 I/S Journal of Law & Policy for the Information Society 543 at 564; Robert H Sloan & Richard Warner, "Beyond Notice and Choice: Privacy, Norms, and Consent" (2013) 14:2 Journal of High Technology Law 370.

[75] Indeed, the drafters of *PIPEDA* thought it "safest to let individuals decide what is sensitive and in which circumstances by giving them control of the information based on the right of consent." See: Office of the Privacy

Is it possible that transparency with respect to automated decision-making will fail to perform instrumentally on account of a similar batch of problems that hinder the efficacy of notice and consent? It seems all too likely given that the general public is *en masse* untrained in applied statistics and related technical fields,[76] limited in the amount of time they're willing to spend interacting with explanatory interfaces, and confused about how to challenge an automated decision or instance of profiling. Despite being painfully obvious, it's too often ignored that the type of information that would be considered 'meaningful' is going to vary among different cohorts of the population. If this epistemic complexity is ignored and the content of 'meaningful information' is created with a prototypically motivated, rational, and technically literate individual in mind, it may be incapable of being understood or used to generate accountability in far too many cases. Simply put, transparency is of no value if what it allows us to see cannot be understood. Proper tools for accumulating and interacting with that information are essential for turning the data into meaningful information to a diverse audience. Specific recommendations for this point are developed in part five.

4.2 : Secrecy-Promoting Concerns

We might also find that our desire for transparency with respect to automated decision-making and profiling conflicts with our competing interests in preventing gaming, safeguarding confidential information and privacy, and deterring stigmatization. Every time that a decision-maker cites one or more of these secrecy-promoting concerns as a justification for withholding

---

Commissioner of Canada, *Consent and privacy: A discussion paper exploring potential enhancements to consent under the Personal Information Protection and Electronic Documents Act* (Report) (Ottawa: Office of the Privacy Commissioner of Canada, 2017) at 2.

[76] Jenna Burrell has identified how technical illiteracy is a driving force of algorithmic opacity. See: "How the Machine 'Thinks:' Understanding Opacity in Machine-learning Algorithms" (2015) 3:1 Big Data and Society 1 at 14.

information about their automated systems, individuals become less and less able to understand and challenge an automated decision or instance of profiling. Therefore, when relied upon to withhold information, secrecy-promoting concerns can undermine the value of transparency.

4.2.1 : Gaming

In analog and automated settings, decision-makers tend to be concerned about individuals gaming their system. Gaming is a threat borne out of the reality that knowledge of the criteria or proxies and their particular weighing that correlate with a target attribute could allow individuals to manipulate those criteria in a calculated effort to improve their performance or otherwise circumvent the purpose of the model. For instance, Kroll et al. write that "if the public knows exactly which items on a tax return are treated as telltale signs of fraud, tax cheats may adjust their behavior and the signs may lose their predictive value for the agency."[77] If decision-makers were to allege that exposing the logic of an automated decision-making system would enable individuals to game and therefore subvert the utility of their system, they could be reasonably reluctant to disclose such information.[78] Therefore, where gaming is a perceived risk, the amount and quality of information that individuals would have to examine would be reduced and their efforts to understand or challenge the legality of an automated decision or instance of profiling would be curtailed.

There are however, two reasons why the automated relative to analog nature of decision-making and profiling might temper how significant a threat gaming actually is to organizations

[77] Joshua Kroll et al, "Accountable Algorithms" (2017) 165 University of Pennsylvania Law Review 633 at 658.
[78] Gaming doesn't always subvert the utility of a decision-making model. There are also instances where gaming can result in socially desirable outcomes and should be encouraged. For example, if a system for measuring creditworthiness takes into account variables that are truly relevant to the outcome that it's measuring, gaming should result in fiscally responsible behaviour and actually increase rather than decrease efficiency. Needless to say, if gaming bolsters rather than undermines the purpose of the automated system it cannot be relied upon as a justification for limiting transparency.

who employ these systems. For one, the non-intuitive, and multi-dimensional nature of certain automated systems utilizing machine-learning techniques might render gaming either impossible or not worth the effort that it takes to achieve it. Second, Michael Brückner et al. and others are working on a technical solution to the problem of gaming called "adversarial prediction games" whereby designing dynamically changing classifiers they can make it substantially more difficult for antagonists to game a model.[79] Where either of these points hold true, it may be that automating decisions actually decreases the risk of gaming relative to human made decisions. As such, I am not convinced that gaming presents such a huge threat to organizations employing automated systems that it will routinely justify them scaling back access to information. Nonetheless, the reality that some models can be gamed and therefore that some information must be kept secret limits the instrumental function that transparency would in principle achieve.

4.2.2 : Confidential Information

Second, an individual's desire to inspect the logic of an automated decision may conflict with a business's interest in protecting confidential information in their source code, algorithms, or decision rules. One could push back against this argument, challenging the efficacy of intellectual property in general and claim that fully open source innovation is a workable business strategy.[80] However, given the lack of mainstream policy support for such a proposal and the general uncertainty surrounding the scalability of open source innovation in the private sector, it is not clear that we should attempt such an overhaul of our treaty-based intellectual property regime

[79] Michael Brückner et al. "Static prediction games for adversarial learning problems" (2012) 13 Journal of Machine-learning Research 2617; Hong Wang et al, "Adversarial prediction games for multivariate losses" (2015) 2 NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems 2728.
[80] Paul de Laat, "Algorithmic Decision-Making Based on Machine-learning from Big Data: Can Transparency Restore Accountability?" (2017) Philosophy and Technology 1 at 12.

for the sake of increased transparency of automated decision-making and profiling. So, for now, to the extent that proprietary interests justify a decision-maker's choice to withhold important information from individuals that would otherwise allow them to understand or challenge a decision, transparency's instrumental value is correspondingly blunted.

4.2.3 : Privacy

Third, granting access to information about the logic involved in automated decision-making or profiling carries a risk of revealing the personal information of others. After all, data mining and machine-learning models are just as reliant on the personal information of the group than the personal information of the individual. In fact a study by Mislove et al. revealed that "multiple attributes can be inferred globally when as few as 20% of the users reveal their attribute information."[81] The tension between transparency and privacy might exist where an individual suspects that an automated decision or profiling tactic is discriminatory and requests access to the data set that the supervised machine-learning algorithm was trained on as well as information about how her result compares to that of the group. Even if efforts were taken to anonymize that data,[82] the persistent prospect of re-identification may raise privacy concerns.[83] Decision-makers might therefore find themselves caught between a rock and a hard place not wanting to be penalized for exposing others' information while simultaneously not wanting to be penalized for failing to provide the individual with meaningful information. To the extent that decision-makers do restrict

[81] Alan Mislove et al, "You Are Who You Know: Inferring User Profiles in Online Social Networks" (2010) Proceeding WSDM '10 Proceedings of the third ACM international conference on Web search and data mining 251.
[82] Through, for example, computational techniques like differential privacy. See: See: Cynthia Dwork et al, "Fairness Through Awareness" (2012) ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 214.
[83] For a legal discussion of the efficacy of different pseudonymization and anonymization techniques, see: Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques 2014/0829.

the amount and quality of information they provide to an individual to protect the personal information of others, individuals lose an opportunity to understand and challenge the automated decision more generally.

4.2.4 Non Stigmatization

Fourth and finally, is the potential concern over transparency and stigmatization raised by Tal Zarsky. Zarsky worries that if access is granted to the correlations and facts of an automated system, "the public, who in its majority is untrained in understanding the intricacies of statistical inferences, will rely upon the information published to reach unfair and wrong social conclusions."[84] There are two arguments that could be made in support of limiting transparency to prevent stigmatization. The first concerns the possibility that based upon the information revealed by the decision-maker, the public interprets group level correlations as reliable signals about individuals within that group. In other words, rather than interpreting a correlation that "people from postal code 'X' are more likely to default on their loan", they will wrongfully deduce that "John from postal code "X" is unreliable and fiscally irresponsible."[85] This is despite the fact that we know that the variance within groups can be greater than the variance between groups and that population level averages are not always precise predictors of traits about individuals within that group. To prevent these sorts of faulty inferences which risk, albeit mistakenly, inflaming stigmatization, one might want to opt for secrecy over transparency. The second argument in favour of justifying secrecy over transparency due to the threat of stigmatization is actually unique to the automated decision-making and profiling context. Namely, while transparency of automated systems may exacerbate the existing problem of stigmatization that results from our collective

[84] Tal Zarsky, "Transparency Predictions" (2013) 4 Illinois Law Review 1503 at 1560.
[85] Thought experiment/structure borrowed from Zarsky, "Transparency Predictions", *supra* note 84 at 156.

problem of statistical clumsiness, it's possible, as Zarsky notes, that data mining may identify new subgroups for stigmatization that do not fall along traditionally well-recognized lines like race, gender, or nationality.[86] As other have noted,[87] discrimination law's history of drawing lines around protected groups will be challenged by data mining's ability to identify currently undefined strata of society based upon shared commonalities that might not be easily definable, nor easily recognizable, due to their seemingly random and non-intuitive nature.[88] However, this unique quality of data mining might be at once that cause and the cure of stigmatization. The more discreet or undefinable these subgroups are, the more inoculated they will be from stereotyping and stigmatization. Beyond just this, there are two other reasons why the prospect of transparency coming at the cost of stigmatization should not justify secrecy over transparency. First, it's not clear that keeping this information secret would prevent or even diminish the prevalence of stigmatization. If anything, more transparency can help push us past our statistical clumsiness, where we foolheartedly and sometimes bigotedly mistake group differences as evidence for characteristics about individuals within that group. Second, even if clawing back information that could be used to stigmatize allayed our concerns about stigmatization, it would probably only do so in a very minor way. In my view, the small benefits obtained by secrecy here do not override the larger instrumental value transparency would otherwise offer.

4.3 : Characteristics of Machine-Learning Techniques

[86] Zarsky, "Transparent Predictions", *supra* note 84 at 1562.
[87] Solon Barocas & Andrew Selbst, "Big Data's Disparate Impact" (2016) 104 California Law Review 671; Toon Calders & Indrė Žliobaitė1, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures" in Custers B et al, eds *Discrimination and Privacy in the Information Society* (Berlin: Springer Berlin Heidelberg, 2014).
[88] Anton Vedder & Laurens Naudts "Accountability for the use of algorithms in a big data environment" (2017) 31:2 International Review of Law, Computers, and Technology 206 at 210.

Similar to how the efficacy of transparency is predicated on a faulty assumption of individuals being universally rational, motivated, and epistemically homogenous, it is also predicated on an untenable understanding of algorithmic systems. Transparency produces best results in environments where decisions are replicable, rule sets are stable, and the reasons for a decision follow a reasonably causal narrative. Automated decision-making and profiling systems which incorporate machine-learning techniques however, can be random, dynamic, non-intuitive, and multi-dimensional. It may be the case that when one or any number of these characteristics are present, generating information about an automated decision or instance of profiling that can be understood or used to generate accountability will be near impossible.

4.3.1 : Randomness

First, while it is true that automated decision-making or profiling systems – including ones employing machine-learning techniques – are for the most part deterministic, they will often incorporate a degree of randomness.[89] While we might also worry about human decision-makers violating legal principles while making random decisions, we might have extra reason to be concerned about automated decision-making systems because here, randomness can be a constituent part of the program. For instance, as already noted, Brückner et al. identify how randomness helps counteract strategic behaviour, or "gaming" of the system.[90] And as Kroll et al. describe, randomness can be used in a decision policy to apportion a scarce resource to equally deserving and qualified candidates in a lottery situation.[91] In other cases, randomness is an essential part of improving the efficiency of a system. Engineers developing machine-learning

[89] Selbst & Powles, "Meaningful Information and the Right to Explanation", *supra* note 64.
[90] Brückner et al, "Static prediction games for adversarial learning problems", *supra* note 79.
[91] Kroll et al, "Accountable Algorithms", *supra* note 77 at 654.

systems must incorporate randomness in order to generalize their models' accuracy beyond their training sets and into new environments, referred to in the technical literature as avoiding overfitting.[92]

The randomness of an automated decision-making or profiling systems presents a threat to an individual's ability to understand and challenge an automated decision or instance of profiling. In the former, randomness poses a threat to an individual's ability to use understanding as a mechanism for creating mental models of the world that can be used for prediction and control. For example, if an individual had access to information like knowing what decision rules the model relied on or what features played most heavily into a decision at point 'A', she could have no confidence that those same rules or features would govern her interactions with that system at point 'B', 'C', and so on. In the latter, randomness may undermine an individual's ability to challenge an automated decision if, for example, the system's randomness obscured the reasons for a decision such that the individual cannot make out whether the decision was inaccurate or unjust.

4.3.2 : Dynamic

Second, and related to the problem of randomness, is that transparency presumes that the model that produced a decision in the past at point "A" is the same as the model that is currently working at point "B". In reality though, machine-learning models can be dynamic; the same set of inputs that produced decision "A" could produce a different decision at point "B". For example, Google updates its search algorithms 500 to 600 times a year.[93] That being the case, any disclosure might be obsolete as soon as its made. If it is not the engineers' tinkering that results in dynamic

[92] Pedro Domingos "A Few Useful Things to Know about Machine-learning" (2012) 55:10 Communications of the ACM 78.
[93] Ryan Shelley, "3 things to do after a major Google algorithm update", *Search Engine Land* (October 18, 2016) online: https://searchengineland.com/3-things-major-google-algorithm-update-260828.

systems, it's the users whose interactions with machine-learning algorithms cause these systems to constantly adapt their behaviour to new inputs. Compounding both these problems is that decision-making and profiling systems are often composed of an ensemble of dynamic algorithms so there is rarely ever one single algorithm to probe. In light of these challenges, we must recognize that transparency with respect to a specific instance of algorithmic activity may only give us a "particular snapshot of its functionality."[94] Like randomness, the intrinsically dynamic nature of some automated decision-making and profiling systems can present a threat to an individual's ability to access and understand information about the logic of that system in order to anticipate how her future interactions with that system are likely to play out. And depending on the speed at which an automated decision-making system is evolving, the fact that it's dynamic may or may not pose a challenge to an individual challenging that decision. For instance, if the system is changing daily or multiple times a day, an individual looking at the information rendered about a decision won't have a solid comparator group to contrast her result with. In this case, a dynamic system might pose an obstacle to an individual understanding whether her decision was unfair or illegal. On the other hand, if the system changes over longer intervals of time, individuals will be in a better position to compare their results to others, understand where their result sits, and judge on that basis whether the decision likely violated any legal rules.

4.3.3 : Inscrutability: Non-Intuitiveness and Multi-dimensionality

Third, transparency can only be of instrumental or intrinsic value if the information it renders is scrutable. However, the machine-learning techniques which frequently underlie these automated systems are optimized for performance and efficiency not narrative coherence or

[94] Mike Ananny & Kate Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability" (2018) 20:3 New Media and Society 973 at 982.

causality. In fact, one of the greatest trade-offs among machine-learning techniques like neural-nets and ensemble methods is that between predictive performance and complexity.[95] As such, any time an automated decision-making or profiling system incorporates machine-learning techniques, it may be impossible or even inefficient to try and deliver an individual with meaningful information about it. Broadly speaking, we can think about the inscrutability of automated decision-making and profiling systems along two dimensions. The first is with respect to the apparent strangeness of a correlation that an algorithmic system may rely on to inform its decision – call this the non-intuitive problem.[96] In the mild case, a perfectly valid automated decision may rely on indicators that we find strange, or non-intuitive. For example, when auditing a spam filter to extract the features most relevant to a classification of 'spam', you might anticipate that words like Nigerian Prince would be highly correlated with positive findings. Instead, Jenna Burrell found that words like 'our' (0.500810), 'click' (0.464474), and 'remov' (0.417698) are better indicators of spam.[97] The more bewildering cases however, as Barocas explains, are those where data mining does "not privilege – or even consider – criteria that hold social salience" but instead relies on "cues to which humans are not physiologically attuned, either because they are very subtle or because they actually fall outside the bounds of humans' perceptual abilities."[98] The non-intuitive feature of machine-learning systems suggests that it might actually be impossible to generate meaningful information in the first place. Clearly, this presents a challenge to an

[95] See: Hoa Khanh Dam et al, "Explainable Software Analytics" (2018) Presented at ICSE'18 NIER online at: https://arxiv.org/pdf/1802.00603.pdf; Tal Zarsky,"The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision-making" (2016) 41:1 Science, Technology, and Human Values 118.
[96] Andrew Selbst & Solon Barocas, "Regulating Inscrutable Systems" (2017) draft paper for We Robot Conference Proceedings, UNIVERSITY OF MIAMI, http://www.werobot2017.com/wp-content/uploads/2017/03/Selbst-and-Barocas-Regulating-Inscrutable-Systems-1.pdf; Andrew Selbst & Solon Barocas "The Intuitive Appeal of Explainable Machines" (2018) 87:XX Forthcoming in Fordham Law Review.
[97] Burrell, "How the Machine 'Thinks'", *supra* note 76 at 8.
[98] Barocas, "Panic Inducing", *supra* note 20 at 119, 122.

individual's ability to understand or challenge such a system. The second dimension is that an automated decision-making system might not just rely on bizarre and non-intuitive correlations; it may rely on hundreds or even thousands of them – call this the multi-dimensionality problem. A computer's unmatched ability to navigate multi-dimensional space is at once a blessing and a curse. On the one hand, their superior capacity to find patterns in massive sets of data has led to expanded access to credit in historically underserved socio-economic segments of the population.[99] On the other hand, as Fayyad points out, "humans are by nature and history dwellers in low-dimensional environments. Our senses and instincts help us deal with three to five dimensions, perhaps as many as 10 if we count all our natural senses and their derivatives."[100] So if an automated decision was made on the basis of 100 or 1000 commingling factors, it may be impossible to understand it given even the most sophisticated data visualization techniques. Together, the non-intuitive and multi-dimensional properties of machine-learning algorithms present a real challenge to providing individuals with access to meaningful information about an automated decision-making or profiling system. Without meaningful information, individuals' efforts to understand and challenge that decision or profiling system fall apart.

Concluding Remarks on the Limits of Transparency

Given these limits, it appears that transparency, as its been described in part two and in the GDPR, is only capable of permitting an individual to understand or challenge an automated decision or instance of profiling if the following conditions hold: the information provided about the automated decision or instance of profiling is presented in a manner that takes into account the

[99] White House Report, "Big Data: A Report on Algorithmic Systems, Opportunity and Civil Rights" (2016) online: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf at 11.
[100] Fayyad, "The Digital Physics of Data Mining", *supra* note 18 at 64.

limitations on individuals' time, rationality, and epistemic variability; there is no risk that revealing the logic involved about the automated system would enable individuals to game and therefore subvert the purpose of that system; revealing information about the decision or profiling would not expose confidential nor others' personal information; and the non-intuitiveness and or multi-dimensionality of that automated decision-making or profiling system doesn't undermine attempts to generate meaningful information in the first place. While there may be a narrow band of cases where transparency in the form of meaningful information about the logic involved in an automated decision or instance of profiling is instrumentally useful pursuant to the GDPR, there are probably a greater number of cases where it will fall short.

Part 5: Recommendations for Supplementing Transparency

Given the limits of transparency so described, Canadian regulators looking to regulate in this space would be ill-advised to duplicate the GDPR's assumption that transparency, in the form of a right to access information, is capable of allowing an individual to understand or challenge an automated decision or instance of profiling. Not all is lost though. Rather, understanding these limits gives us the opportunity to craft a regulatory framework that considers the complex interaction between them. To that end, this section advances two sets of recommendations. First, where information about automated decision-making or profiling is capable of being produced without raising secrecy-promoting concerns or being undermined by the characteristics of machine-learning techniques, that information must be delivered in a variety of ways in order to increase the chance that it will be meaningful to an epistemically diverse audience. Second, where secrecy-promoting concerns do prevent individuals from accessing this information or where the characteristics of machine-learning conspire to render this information meaningless, organizations

must have their automated decision-making and profiling systems verified by a neutral third party. I float the idea of creating a certification body run by OPC to act as the neutral third party to fulfill this role.

5.1 : Accounting for Epistemic Complexity in Meaningful Information

The problem at the core of the epistemic complexity critique with respect to automated decision-making and profiling is that the type of information that would be delivered under the banner of 'meaningful information' is likely to be meaningful – in the instrumental sense – to a small subset of people. Collapsing epistemic complexity down into an amorphous standard of the reasonable person has long been a specialty of legal doctrine, such as in the case of what qualifies as libel.[101] In the context of automated decision-making, it's therefore no surprise to come across proposals by the likes of Reuben Binns who has suggested that "the notion of *public reason*—roughly, the idea that rules, institutions and decisions need to be justifiable by common principles, rather than hinging on controversial propositions which citizens might reasonably reject—is an answer to the problem of reasonable pluralism in the context of algorithmic decision-making [my emphasis]."[102] However, we might want to ask whether we should move past the reasonable person standard and account for epistemic complexity in this context. I see two reasons for doing so.

First, the greatest barrier to accounting for epistemic complexity – cost – is reduced by orders of magnitude in a computational setting relative to its analog counterpart. While some costs are bound to exist in the automated case, they do not include the cost of having a human respond to individual requests for access to information at a level of complexity that the individual seeking

[101] *Colour Your World Corp v Canadian Broadcasting Corp* (1998), 156 DLR (4th) 27 (Ont CA), per Abella JA at para 36.
[102] Reuben Binns, "Algorithmic Accountability and Public Reason" (2017) Philosophy and Technology 1 at 3.

that information would consider meaningful. Rather, user interfaces can be created which allow individuals to seamlessly access information that they would subjectively consider meaningful. And once these user interfaces are up and running, the cost of maintaining them should drop precipitously.

The second reason for embracing this recommendation is because in general, we expect the quality of an explanation to scale with the level of risk a decision presents us with.[103] That's why we expect hundred-page long explanations from the Supreme Court and "because you also watched" type explanations for why we're recommended a movie on Netflix. Because the systems that ought to be included in this regulatory regime are those which produce legal and similarly significant levels of risk it is reasonable to demand access to a wider variety of meaningful information. To be meaningful, this information should be capable of allowing individuals of various ages, levels of ability and literacy, and knowledge of avenues of institutional recourse (i.e. how to file a complaint) to understand and if necessary act upon that information. Organizations should therefore be required to employ a wide range of technical tools and design strategies to produce information which caters to a similarly wide range of individuals' epistemic needs such that it is instrumentally useful.

To reiterate, because the cost of delivering meaningful information to a plurality of epistemic needs is significantly reduced by its amenability to being automated, the main barrier to delivering information in this way is knocked down considerably. Moreover, because the quality of an explanation should scale with the level of risk a decision presents us with and the type of automated decision-making and profiling being regulated is that which presents individuals with legal or similarly significant risks, we are entitled to expect organizations using these systems to

[103] I don't mean to insinuate that an explanation be the only type of information an individual has access to.

deliver correspondingly high-quality information. Considering how important delivering subjectively meaningful information to the widest possible audience is to the project of transparency, I recommend that it form a core pillar of a Canadian regulatory strategy.

To overcome this epistemic variability and the limitations of the prototypically motivated, rational, technically literate, and informed individual, these interfaces should nudge individuals into engaging with them. Without this, information is likely to be ignored or worse, instill a false sense of security among individuals wherein an organization's mere provision of information masquerades as compliance with legal rules. To avoid both of these outcomes, meaningful information can be presented in a combination of graphical or textual forms and employ audio-visual tools to support the digestion of that information.[104] That information should also be generated at a number of levels of complexity having due regard for variability in age, technical and legal literacy, time, and knowledge about avenues of institutional recourse. So, whereas a lengthy technical description may be exactly what a data scientist and a lawyer need to challenge an organization, a graph and a short explanation of how car insurance applicant 'A' compares with others similarly situated to him may be suitable for a young adult. If that graph and explanation reveals him to be an outlier, he should be able to navigate that same explanatory interface to access increasingly granular information that would facilitate his understanding and ability to challenge the decision. Understandably, there is bound to be some looseness of fit between an explanatory interface and the epistemic idiosyncrasies of the particular individual seeking such information. The goal however, ought to be to reduce this gap to the greatest degree possible. Given how low

[104] To give the GDPR credit, we do find in Article 12 a requirement that "any information to be communicated the data subject in a "concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child." Moreover, Recital 58 and the Article 29 WP Guidance on Automated Decision-making elaborate on this point suggesting that where appropriate, visualization and interactive techniques to aid algorithmic transparency *may be used*. These techniques *must* be used though. GDPR, *supra* note 5 at Recital 28; Article 29 WP, *supra* note 35 at 28.

the bar is set currently, there is room for rapid and marked improvements. To get here, bridges will need to be built between organizations employing these systems and experts in the fields of human computer interaction and user experience design,[105] data visualization,[106] and the social and computer sciences.[107] In short, we want organizations to make as creative a use of their platforms to facilitate transparency as the way their platforms make use of us and our data.

5.2 : Neutral Third Party: Certification Body

As described in part four, the collection of secrecy-promoting concerns – gaming, confidential information, and privacy (but not stigmatization) – will at times limit the amount and meaningfulness of the information that organizations can provide to individuals about their automated decision-making and profiling systems. However, as soon as we realize that secrecy-promoting concerns are really only *concerns* when information is disclosed to the wrong sets of eyes – that is individuals themselves – methods of averting these concerns surface. In particular, by shifting the recipient of transparency and meaningful information to a neutral third party, we should find that these secrecy-promoting concerns largely fall away. If gaming was the issue, the organization would disclose the features that factor most heavily into their model's performance so the neutral third party could inspect whether the organization's model was redundantly encoding a prohibited ground of discrimination. If confidential information was the concern, the

---

[105] Lachlan Urquhart & Tom Rodden, "New directions in information technology law: learning from human–computer interaction" (2017) 31:2 International Review of Law, Computers, and Technology 150; Michael Veale & Reuben Binns Max Van Kleek, "Some HCI Priorities for GDPR-Compliant Machine-learning" (2018) Workshop at ACM CHI'18.

[106] Edward Tufte, *The Visual Display of Quantitative Information*, 2nd ed (Connecticut: Graphics Press, 2001); Laurence Diver & Burkhard Schafer, "Opening the Black Box: Petri nets and Privacy by Design" (2017) 31 International Review of Law Computers and Technology 68.

[107] Tim Miller "Explanation in Artificial Intelligence: Insights from the Social Sciences" (2017) online: https://arxiv.org/abs/1706.07269; Wachter et al, "Counterfactual Explanations Without Opening the Black Box", *supra* note 70; Dam et al, "Explainable Software Analytics", *supra* note 95.

organization would disclose the source code or algorithms to the neutral third party. This would enable black or white box testing on the model to find out for example its false positive or negative rate and whether misclassifications are unevenly distributed across the population.[108] And if a concern about revealing personal information was the limiting factor, the organization would provide the neutral third party with the data used to train the decision-making or profiling system to inspect whether those data are accurate[109] or proportionately representative of the population for which they will be used to make a decision about.[110] In every case, when delivering direct and individualized access to information about a solely automated decision-making or profiling system producing legal or similarly significant effects would pose a real and substantial threat to the organization's confidential information or the personal information of others, they should have to instead submit that information to a neutral third party for review *before* they put their decision-making or profiling system into use.[111] Based upon this *ex-ante* analysis of the decision-making or

---

[108] For a discussion of black and white box testing, see: Kroll et al, "Accountable Algorithms", *supra* note 77.

[109] On this last point, an FTC study found 21% of its sample of consumers had confirmed an error on at least one of their three credit reports: White House Report, "Big Data: A Report on Algorithmic Systems, Opportunity and Civil Rights" (2016) online: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf; When a data set is proportionately representative enough to survive scrutiny is something I do not have an answer to.

[110] The Supreme Court of Canada in *Ewert v Canada* recently noted (in the context of algorithmically generated risk recidivism assessments) that tools developed and validated based on data from majority groups may lack validity in predicting the same traits in minority groups. *Ewert v Canada*, [2018] SCC 30 at paras 13, 41.

[111] While such a determination (e.g. whether an organization's use of k-anonymity to anonymize personal information means that they should deliver that anonymized information to the public or not) is likely to be somewhat context specific, from the standpoint of administrability, a threshold legal rule is necessary. As identified, this test could take the following shape: "*delivering direct and individualized access to information about a solely automated decision-making or profiling system producing legal or similarly significant effects would pose a real and substantial threat to the organization's confidential information or the personal information of others*." Breaking this down further, the certification body must ascertain whether the information the organization wants to withhold from individuals is in fact confidential or personally identifiable information. While a full discussion of confidential information and privacy law on these matters exceeds the scope of this discussion, the following parameters ought to inform the certification bodies analysis. Pursuant to case law such as *Merck Frosst Canada Ltd v Canada (Health)*, 2012 SCC 3, and *Lac Minerals Ltd v International Corona Resources Ltd*, [1989] 2 SCR 574, whether information is confidential will turn on questions of whether that information is treated as a secret in an absolute or relative sense (i.e. known by a relatively small number of persons and measures are taken to keep it secret); is capable of industrial or commercial application; and is worthy of legal protection due to the economic interest of the possessor. A similar definition of confidential information is found in Article 39 paras 1-3 of *The Agreement on Trade-Related Aspects of Intellectual Property Rights* (TRIPS) to which Canada is a party. Here,

profiling system, the neutral third party could either approve or disapprove of it being used among the general public.[112] A stamp of approval or certification mark from this third party would effectively vouch to individuals that they can trust that the decision-making or profiling system is acting in accordance with the law without compromising the interests in that sensitive information. And despite the apparent inability of privacy seals to achieve a similar set of stated objectives, we may avoid a similar fate here considering that these marks will be generated directly by the OPC for automated activities that may pose a higher – legal or similarly significant – level of risk.

The same recommendation applies with respect to overcoming the limits imposed by the characteristics of machine-learning. That is, a neutral third party could house a decision policy locked in through technological measures like cryptographic guarantees and zero-knowledge proofs to verify that despite a decision or profiling system's inscrutability, the organization responsible for it is verifiably acting within the bounds of the law. Joshua Kroll describes cryptographic commitments as a method that "can be used to lock in knowledge of a secret (say, an undisclosed decision policy) at a certain time (say, by publishing it or sending it to an oversight

---

information may be considered confidential so long as (a) is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question; (b) has commercial value because it is secret; and (c) has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret. See Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr 15, 1994, 1869 UNTS 299, 33 ILM 1125, 1197. Pursuant to the OPC and case law, personal information is information about an identifiable individual. An individual may be considered 'identifiable' if there is a serious possibility that he or she could be identified through the use of that information, alone or in combination with other information. See *Gordon v Canada (Health)*, 2008 FC 258. Ensuring that there is a real and substantial risk of revealing confidential or personal information is necessary to ensure that organizations do not move to have all of their activities verified by the certification body rather than providing individuals with this information directly. Such an outcome is undesirable for at least two reasons. For one, every time these concerns are justified, individuals lose the ability to understand and challenge these decisions directly. Substituting verification for direct understanding and accountability is a compromise that ought to be justified only where a real and substantial threat can be found. Second, if verification becomes the rule rather than the exception, there is a possibility that the certification body might succumb to industry capture given that their primary and repeat business would be from organizations rather than the individuals directly and significantly affected by such decisions.

[112] Because automated systems are generally deterministic in their outputs, these *ex-ante* verifications should ensure their tri-annual validity. For a discussion of the determinism of machine-learning, see: Selbst and Powels, "Meaningful Information and the Right to Explanation," *supra* note 63 at 16.

body) without revealing the contents of the secret, while still allowing the secret to be disclosed later (e.g., in a court case under a discovery order) and guaranteeing that the secret was not changed in the interim (for example, that the decision policy was not modified from one that was explicitly discriminatory to one that was neutral)."[113] Zero-knowledge proofs on the other hand are useful for ensuring that the rules published by the controller were actually the rules used to make a given decision.[114]

Technological solutions are not the only methods of achieving accountability in the face of technologically driven opacity though. Solon Barocas and Andrew Selbst have suggested that in the face of opaque and non-intuitive algorithmic systems, organizations should be prepared to present "documentation" of the "institutional and subjective process" behind the automated system's development.[115] This information would provide a neutral third party with another angle to inspect whether the inscrutable decision-making or profiling system is acting in concert with "society's broader normative priorities, as expressed in law and policy."[116]

A short discussion of feature engineering, class specification, and algorithmic selection will help bear out how documentation might be relevant to our normative priorities. Feature engineering is a stage in the machine-learning process that involves synthesizing data into features which might be relevant for solving a problem.[117] It's a finicky process wherein features that are too wed to the particularities of the training data will lead to a problem called overfitting, where the model becomes loses predictive performance when fed new data.[118] But if too many features are created, analysts run into the "curse of dimensionality" which refers to the finding that the

[113] Kroll et al., "Accountable Algorithms," *supra* note 76 at 666.
[114] Kroll et al., "Accountable Algorithms," *supra* note 76 at 668.
[115] Barocas and Selbst, "Intuitive Appeal of Explainable Machines", *supra* note 96 at 55.
[116] Barocas and Selbst, "Intuitive Appeal of Explainable Machines", *supra* note 96 at 59.
[117] Domingos "A Few Useful Things to Know about Machine-learning", *supra* note 92.
[118] Domingos "A Few Useful Things to Know about Machine-learning", *supra* note 92.

amount of data you need grows exponentially with the number of features you create.[119] An example of feature engineering might be where a programmer selects certain properties of an e-mail — the words in the body of the e-mail, the names of the attachments, as well as the words in the subject line — and leaves it to the algorithm to figure out which words are spammy and how spammy they are.[120] Or as Veale and Binns explain, feature engineering could involve aggregating those who subscribe to different branches of a religious doctrine (e.g. Catholic, Protestant; Shia, Sunni) within a single overarching doctrine (Christian, Muslim).[121] The risk here is that that the feature engineering process might collapse distinctions by paying more or less attention to certain data which can be highly relevant to questions of fairness and discrimination.[122] Access to how features were constructed might be useless to the layperson, but may be highly relevant to a certification body's assessment of an automated system's fairness.

Second, the specification of *class definitions* (or in statistical terms target variables) in supervised learning systems significantly impacts how that system will perform and may be relevant to the verification of a technically inscrutable system. Here, analysts describe what it means for something to be a case of spam, for someone to be at risk for recidivism, or for someone to be a good credit risk.[123] Obviously, specifying certain class definitions will be a binary or objective task as in the automated classification of a cat or a dog. Other cases however, such as the determination of creditworthiness or employability are more ambiguous and reasonable people can debate the boundaries of these classes. If the public was given access to debate these class definitions though, their access may come at the cost revealing the organization's confidential

[119] Domingos "A Few Useful Things to Know about Machine-learning", *supra* note 92.
[120] Andrew Tutt, "An FDA for Algorithms" (2017) 69:1 Administrative Law Review 83 at 96 - 97.
[121] Veale & Binns, "Fairer ML in the real world", *supra* note 29 at 2.
[122] Veale & Binns, "Fairer ML in the real world", *supra* note 29 at 2.
[123] Domingos "A Few Useful Things to Know about Machine-learning", *supra* note 92.

information. Moreover, these boundaries may also be of such a technical nature that they may lead the public to draw faulty conclusions. A certification body though, need not be hampered by either of these concerns. In principle, it could examine that information without compromising an organization's interests in keeping it away from prying eyes while also having the technical literacy to do so effectively.

Third, the selection of one algorithm over another might be relevant to our normative aims. As a policy matter, the choice of one method over another may be relevant in a high stakes decision where an organization chooses an inherently opaque and non-intuitive method over a more transparent one where the benefits of the more opaque method are insignificant.[124] Even when an organization can document that gains in efficiency justify the selection of less interpretable methods, it should still be taken into account whether state of the art technical methods were used to salvage the scrutability of the models and to provide individuals with direct access to information about them.[125]

This is obviously not an exhaustive list of technical solutions to algorithmic opacity or of the type of information we might want organizations to provide a certification body with under the banner of 'documentation'.[126] Rather than being prescriptive of what documentation ought to look like, being flexible here would preserve a general commitment to technological neutrality and would appreciate that different contexts are likely going to require different technological methods and types of documentation. Such flexibility though, need not be paralyzing from an

[124] Where the desire for transparency trumps marginal gains in efficiency we might be more skeptical of generally unintelligible methods like boosted trees, random forests, bagged trees, kernelized-SVMs, neural nets and deep neural nets as compared to more interpretable methods like decision trees, naïve Bayes classification, and rule learners.

[125] Riccardo Guidotti et al, "A Survey Of Methods For Explaining Black Box Models" (2018) online: https://arxiv.org/pdf/1802.01933.pdf.

[126] For another discussion, see: Argyro Karanasiou & Dimitris Pinotsis "A study into the layers of automated decision-making: emergent normative and legal aspects of deep learning" (2017) 31:2 International Review of Law, Computers, and Technology 170.

administrability perspective. Rather, an organization's verification can be contingent upon whether the technical guarantees they bake in or documentation they provide facilitates a neutral third party's assessment of whether the automated decision-making or profiling system is operating within the bounds of the law. The value of transparency here is still instrumental, it has just shifted to a different set of eyes.

I am not the first to recommend that a neutral third-party step in to generate accountability over algorithmic systems. Andrew Tutt for example, has floated the idea of establishing an "FDA for algorithms";[127] Ryan Calo has similarly considered the benefits of a "federal robotics commission";[128] and Frank Pasquale has suggested having a "trusted auditor" examine algorithmic systems in order to bypass legitimate, but sometimes inflated, concerns about confidentiality.[129] While the arguments for such proposals largely track the ones discussed here, they are somewhat wed to an American regulatory landscape. That being the case, what would a neutral third party look like in Canada? I suggest that the Office of the Privacy Commissioner (OPC) ought to take up this role. Specifically, the OPC should house a certification body which would be tasked with verifying the automated decision-making and profiling systems of organizations who on account of secrecy-promoting concerns or the characteristics of machine-learning techniques cannot be transparent to individuals directly. The OPC would issue marks or seals to organizations who could demonstrate through technical guarantees, documentation or other means that their automated systems are legally compliant. These seals of certification could last for as long as the automated

[127] Andrew Tutt, "An FDA for Algorithms" (2017) 69:1 Administrative Law Review 83.
[128] Ryan Calo, "The Case for a Federal Robotics Commission" (2014) Brookings Institution Center for Technology Innovation, online: https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission/.
[129] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015) at 141; see also Citron, Danielle & Pasquale, Frank. "The Scored Society: Due Process for Automated Predictions" (2014) 89:1 Washington Law Review 1.

decision-making or profiling system remain substantially equivalent up to a period of three years, upon which they would need to be renewed.[130]

Part 6: Counterarguments and Regulatory Costs

In this section, I want to address two potential counter-arguments that may be laid at the feet of these proposals. In the former, do we risk burdening organizations with the requirement to accommodate a plurality of epistemic needs? In particular, are these costs likely to be disproportionately borne by small to medium size enterprises? In the later, do we really want to empower the government, and in particular the OPC, to take up the task of certifying automated decision-making and profiling systems? Might privately run certification bodies be better equipped to certify organizations who cannot be transparent to individuals directly on account of secrecy- promoting concerns or technical inscrutability? I address each in turn before raising one final critique of certification bodies that deserves further legal consideration.

6.1 : Will generating meaningful information become an unduly costly activity?

Designing sleek and useful user interfaces capable of being considered meaningful to a diverse audience is going to impose additional organizational costs relative to the status-quo. [131]

[130] This standard will allow organizations to make iterative improvements to their automated decision-making and profiling systems without having to have them re-verified each time as long as their systems' specification remains substantially similar. For example, a system may be considered no longer substantially similar if adjustments are made to the decision policy it locked in with a cryptographic proof in the first instance. I acknowledge the room for ambiguity in this standard. Indeed, there is a danger that organization will sneak in large changes to their systems under the guise of them retaining the originally functionality of their verified system. However, I run up against the limits of my technical knowledge when attempting to elucidate when an automated decision-making system will or will not be considered substantially similar. Technical and legal collaboration would be needed to flesh out this standard more fully if Canada moves in this regulatory direction.

[131] To read a critique of the precautionary principle *vis a vis* permissionless innovation, see: Adam Thierer et al, "Artificial Intelligence and Public Policy" (2017) Mercatus Research at George Mason University, available at SSRN: https://ssrn.com/abstract=3021135.

For larger organizations, these costs may not be inordinate. However, we might be concerned about small to medium size enterprises who may not be able to staff departments composed of experts in machine-learning, user experience design, and human-computer interaction. If regulation of this sort were to disproportionately affect smaller companies from competing, it might have the unfortunate knock-on effect of reinforcing the dominance of the so-called 'technoligarchy' (Microsoft, Google, Facebook, Amazon) in the automated decision-making and profiling market. If this requirement is indeed so burdensome, it perhaps even risks fostering a regulatory climate that is inhospitable to industry and may thwart Canada's efforts to compete in the development and implementation of AI.

According to this line of thought, the costs associated with generating meaningful information that takes into account epistemic variability would indeed be significant. What are these costs though? And are they likely to be so great that they trump the normative impetus behind generating information that individuals of different ages, technical abilities, and institutional knowledge would consider meaningful? The upfront costs of generating this information may indeed be large as most organizations will find themselves having to either retrain or hire new staff skilled in the areas of user experience design, human-computer interaction, and data visualization. However, it may be the case that organizations already generate some of this information in the process of creating and testing their own models and it may only be a matter of translating it into more accessible forms. Moreover, after this upfront effort to build user interfaces capable of delivering this information, the work is largely done – especially if off the shelf solutions emerge as viable solutions for certain classes of algorithms.[132] Because transparency delivered directly to

[132] While maintenance will be likely required from time to time and efforts can always be made to improve the quality of the explanatory interfaces, it's not clear that these ongoing efforts would constitute an undue financial strain relative to most organizations' normal operations.

individuals can be largely automated after this point, it should represent a far smaller cost relative to requiring an organization to deliver a human-generated explanation of each decision to every individual. Certainly, this specific requirement is unlikely to undercut Canada's efforts to cultivate a competitive and innovative AI sector. In fact, the inverse might be true if regulating automated decision-making in the private sector ends up being a prerequisite to gaining adequacy standing *vis a vis* the GDPR. In this scenario, leapfrogging the GDPR in this respect would facilitate rather than constrain innovation and economic development by reducing transaction costs associated with handling personal information across jurisdictions. Finally, alongside the Canadian government's commitment to fostering a competitive and innovative technology sector through investments into a Pan-Canadian AI strategy, its desire to emerge on the world stage as a leader with respect to the ethical development and implementation of AI is equally strong.[133] Moving forward with a regulatory regime that defines meaningful information in an inclusive way as recommended here would be a step towards solidifying Canada's position as a leader in this space. Altogether then, while the exact dollar figure associated with doing transparency as discussed here is difficult to pin down, I do not anticipate it to be so great a regulatory burden that represents a real barrier to industry even among small to medium size enterprises.

6.2: Is a Government Run Certification Body the Answer?

Regarding my recommendation that the OPC steps in to verify automated decision-making and profiling systems in the cases where secrecy-promoting concerns or the characteristics of machine-learning make it undesirable or impossible to deliver meaningful information to

[133] For instance, in June, 2018 at the G7 Leaders' Summit, Canada committed to, with the support of France, to head up a working group on the study of the human rights, geopolitical, and democratic implications of AI.

individuals directly, the following two critiques can be made. First, what if the OPC lacks the institutional competence to take on the highly technical role as a certification body for automated decision-making and profiling systems? Second, if the OPC was tasked to play this role, they would be granted access to incredible swaths of Canadians' data which is at risk of being breached, or illegally shared among departments.

Considering these complaints, perhaps there are other more effective ways to allay our concerns when transparency cannot be generated towards individuals directly. For example, the GDPR prescribes a framework where certification bodies would exist as arms-length private entities that would be authorized by local data protection authorities to certify that organizations are compliant with the regulation. Would this approach be any better? We might say that arms-length certification bodies would be capable of attracting better technical talent relative to governments. However, this assumes that certification bodies would indeed be in a financial position to offer more competitive salaries than governments. It's not clear that market forces would necessarily turn privately run certification bodies into a lucrative industry capable of offering these salaries though. Even if we granted that the private sector would be more capable of attracting the human capital necessary to certify other private organizations, there appear to be at least three other reasons to be skeptical of preferring this approach. For one, like governments, the private sector is also vulnerable to data breaches and succumbing to the temptation of misusing the vast swaths of information that would be vested within them. Second, given that the role of the certification body would be to verify organizations' automated decision-making and profiling systems' compliance with the law, its activities have a direct impact on the public interest and therefore ought to be housed in a governmental body operating in accordance with administrative principles. Third, if certification bodies were to be arms-length private organizations, their

certification marks may lose credibility and come to be viewed with similar skepticism or disregard as privacy seals.

Whether a certification body exists within the OPC or at an arms-length distance in the private sector though, there appears to be one outstanding problem. That is, we may find ourselves in a position where we will want to demand transparency with respect to the decisions made by that certification body. But given that the certification body's existence is borne out of a need to counterbalance legitimate desires for secrecy, what kinds of information would we expect a certification body to be transparent about? More specifically, if housed in the OPC, how would the administrative law principle of procedural fairness, which among other things requires administrative decisions to be made in the open, apply to a certification body whose function is to make decisions about confidential and private information?[134] Moreover, given the inherent complexity in an increasing number of automated decision-making systems, certification bodies may need to utilize sophisticated and complex tools to certify that organizations' inscrutable automated systems are operating within the bounds of the law. As noted by Desai and Kroll, "insofar as the techniques are based on machine-learning, the irony may be that the techniques will be as inscrutable as the systems they mean to analyze and thus will fall short of providing technical accountability."[135] Will the next step be creating a public algorithmic review board to scrutinize a certification body that scrutinizes a private company? If these concerns are left unanswered, my recommendation for creating a certification body portends to replicate a similar class of problems that it attempts to overcome. That is, how should we think about transparency in relation to a certification body given epistemic complexity, secrecy-promoting concerns, and the

[134] *Baker v Canada*, *supra* note 6 at para 44.
[135] Joshua Kroll & Deven R. Desai, "Trust but Verify: A Guide to Algorithms and the Law" (2018) 31:1 Harvard Journal of Law and Technology 1 at 18.

characteristics of machine-learning? Are we willing to forfeit some transparency with respect to this certification body to achieve the accountability of automated decision-making and profiling systems? This trade-off wouldn't be entirely unique to automated decision-making and profiling.[136] Moreover, there is a body of administrative law that would govern the relationship between individuals and such a certification body. While beyond the scope of this paper, the relationship between administrative law and my proposal for a certification body ought to be mapped out if Canada is going to move forward with this proposed regulatory framework.[137]

Conclusion

As I have attempted to demonstrate in this paper, a confluence of epistemic variability, secrecy-promoting concerns, and characteristics of machine-learning may undermine the extent to which transparency can achieve its stated aims. That is, transparency, in the form of an individual right to access information, is by itself an insufficient regulatory response to enabling an individual to understand or generate accountability with respect to automated decision-making and profiling systems. To overcome these limits, there are two revisions which can be made. First, individuals must be provided with meaningful information about the rationale and process involved in an automated decision-making or profiling system where 'meaningful information' is defined as that which permits an individual to understand or challenge such a system having *due regard for*

[136] For example, Health Canada protects confidential and personal information during the drug and health product reviewal and approval process. This information may only be disclosed to protect or promote public health or safety. See: Health Canada, *Request for disclosure of confidential business information* (Ottawa: Health Canada, 2018) online: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/request-disclosure-confidential-business-information.html.

[137] In addition to art 39 paras 1-3 of TRIPS, North American Free Trade Agreement Between the Government of Canada, the Government of Mexico and the Government of the United States, 17 December 1992, Can. TS 1994 No 2, 32 ILM 289 art 1711(5) as well as *Access to Information Act*, RSC, 1985, c A-1 s 20(1) lay out conditions where confidential information is protected as a part of a regulatory approval process and the conditions where it will or will not be disclosed e.g. to protect the public or take steps to protect information from unfair commercial use.

variability in age, technical and legal literacy, time, and knowledge about avenues of institutional recourse. However, even with such a provision in place there will be cases where a confluence of secrecy-promoting concerns like gaming, confidential information, and privacy along with the non-intuitive and multi-dimensional properties of machine-learning techniques make it undesirable if not impossible to deliver meaningful information so defined. In these cases, transparency in the form an individual access right to meaningful information promotes neither understanding nor accountability. If, however, the recipient of meaningful information shifts over to a neutral third party like a certification body housed within the OPC, accountability over these automated systems can still be achieved. Here, an organization that's uncomfortable (due to secrecy-promoting concerns) or unable (due to the characteristics of machine-learning techniques) to provide individuals with direct access to meaningful information about the rationale and process involved in their automated decision-making or profiling system could seek verification from the certification body before putting that system into general use. This verification process could hinge on whether the organization is capable of demonstrating that their automated system is legally compliant. To demonstrate compliance, an organization could submit their sensitive information for inspection and demonstrate through documentation as well technical measures that their decision-making or profiling system is operating within the bounds of the law. This particular recommendation is somewhat coarse and the exact details of what a verification process ought to look like is a bit of an open question. Without a doubt, a more granular account of this process ought to be the product of a multi-disciplinary collaboration between technical and legal communities.

As far as the types of automated decision-making and profiling systems that should fall under the scope of this proposed regulatory regime, I am inclined to take inspiration from the

GDPR. That is, individuals should have access to *meaningful information* about the rationale and process involved in *solely* automated decision-making *and* profiling systems which produce *legal* effects or *similarly significantly* affects him or her. Where organizations are unwilling or unable to provide access to meaningful information directly to individuals, their automated decision-making or profiling system must be verified by a certification body tri-annually.

Alongside filling the gaps in applying the logic of transparency to automated decision-making and profiling, this discussion has identified a new set of questions that emerge from the proposed recommendations. My first recommendation raises a challenge for user-experience designers, data-visualization experts, human-computer interaction experts, and policy experts to collaborate in an effort to make technical information comprehensible to the widest possible audience.[138] The larger set of questions however, arise with respect to establishing a certification body in the OPC to pre-emptively certify organizations using automated-decision making or profiling systems who cannot or will choose not to reveal meaningful information directly to individuals on account of secrecy-promoting concerns or the characteristics of machine-learning. For example, under which circumstances will organizations be justified in withholding meaningful information from individuals directly and be permitted to avert the costs of generating such information and have their systems verified by the certification body? Second, given the iterative and dynamic process of designing and maintaining these systems, might a system change so much in a three-year period that its original verification no longer holds? What exactly does this verification process entail? Finally, what type of information can we expect the certification body

[138] For an example of techno-legal collaboration, see: Association for Computing Machinery US Public Policy Council, "Statement on Algorithmic Transparency and Accountability" (2017) online: https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf; IJCAI-17 Workshop on Explainable AI (XAI) Proceedings. (2017) Melbourne, Australia online: http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf.

to disclose to us with respect to how it verifies these organization's automated systems? What lessons can be drawn from regulatory bodies like Health Canada who are in a similar position of having to pre-emptively verify a product where there are legitimate concerns about confidential and personally identifiable information? While I have provided partial answers to these questions, they ought to be debated more widely before such a recommendation can be confidently operationalized.

LEGISLATION

Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr 15, 1994, 1869 UNTS 299, 33 ILM 1125, 1197.

A Local Law (New York City) in relation to automated decision systems used by agencies Online: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

Access to Information Act RSC, 1985, c A-1.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data [1995] OJ L 281/31.

Digital Privacy Act SC 2015, c 32.

Freedom of Information and Protection of Privacy Act, RSO 1990, c F 3.

North American Free Trade Agreement Between the Government of Canada, the Government of Mexico and the Government of the United States, 17 December 1992, Can. TS 1994 No 2, 32 ILM 289

Personal Information Protection and Electronic Documents Act (*PIPEDA*), SC 2000, C 5.

Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Advancement of Such Data, and repealing Directive 95/46/EC, 2016 OJL 119/1.

JURISPRUDENCE

Baker v Canada (Minister of Citizenship and Immigration) [1999] 2 SCR 817.

Canada (Information Commissioner) v Canada (Transportation Accident Investigation and Safety Board), 2006 FCA 157.

Canada (Information Commissioner) v Canada (Commissioner of the Royal Canadian Mounted Police), [2003] 1 SCR 66, 2003 SCC 8.

Case 215/88 Casa Fleischhandels [1989] European Court of Justice ECR 2789.

Colour Your World Corp v Canadian Broadcasting Corp (1998), 156 DLR (4th) 27 (Ont CA).

Dagg v Canada (Minister of Finance), [1997] 2 SCR;

Ewert v Canada, 2018 SCC 30.

Gordon v Canada (Health), 2008 FC 258.

Lac Minerals Ltd v International Corona Resources Ltd, [1989] 2 SCR 574.

Merck Frosst Canada Ltd v Canada (Health), 2012 SCC 3.

SECONDARY MATERIALS

Ananny, Mike & Crawford, Kate. "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability" (2018) 20:3 New Media and Society 973.

Angwin, Julia et al. "Machine Bias", *Pro Publica* (23 May 2016) online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Association for Computing Machinery US Public Policy Council. "Statement on Algorithmic Transparency and Accountability" (2017) online: https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf.

Article 29 Data Protection Working Party, Guidelines on automated individual decision-making and Profiling for the purposes of Regulation 2016/679.

Article 29 Data Protection Working Party, Guidelines on transparency under Regulation 2016/679.

Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques 2014/0829.

Bamberger, Kenneth A & Mulligan, Deirdre. "Privacy Decision-making in Administrative Agencies" (2008) 75:1 Chicago L Rev 75.

Barocas, Solon. "Panic Inducing: Data Mining, Fairness, and Privacy" (PhD dissertation, New York University, 2014).

Barocas, Solon et al. "Governing Algorithms: A Provocation Piece" (2013) Paper prepared for the Governing Algorithms conference, available at SSRN: https://ssrn.com/abstract=2245322 or http://dx.doi.org/10.2139/ssrn.2245322.

Barocas, Solon & Selbst, Andrew. "Big Data's Disparate Impact" (2016) 104 California Law Review 671.

Baratta, Roberto. “Complexity of EU Law in the Domestic Implementing Process” (2014) 2 The Theory and Practice of Legislation 29.

Bayamlıoğlu, Emre. “Transparency of Automated Decisions in the GDPR: An Attempt for systemization” (2018) Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3097653.

Binns, Reuben. “Algorithmic Accountability and Public Reason” (2017) Philosophy and Technology 1.

Birkinshaw, Patrick. “Freedom of Information and Openness: Fundamental Human Rights?” (2006) 58 Administrative Law Review 177 at 189.

Brkan, Maja. “Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond” (2018) submitted for ‘Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence’ A previous version of this paper was published as Brkan, M. (2017), ‘AI-supported decision-making under the General Data Protection Regulation’, in Proceedings of the 16th International Conference on Artificial Intelligence and Law, London.

Brückner, Michael et al. “Static prediction games for adversarial learning problems” (2012) 13 Journal of Machine-learning Research 2617.

Burrell, Jenna. “How the Machine 'Thinks:' Understanding Opacity in Machine-learning Algorithms” (2015) 3:1 Big Data and Society 1.

Calders, Toon & Žliobaitė, Indrė. “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures” in Custers B et al, eds *Discrimination and Privacy in the Information Society* (Berlin: Springer Berlin Heidelberg, 2014).

Calo, Ryan. “The Case for a Federal Robotics Commission” (2014) Brookings Institution Center for Technology Innovation online: https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission/.

Casey, Bryan et al. “Rethinking Explainable Machines: The GDPR’s “Right to Explanation” Debate and the Rise of Algorithmic Audits in Enterprise” (2018) Forthcoming in Berkley Technology Law Journal.

Citron, Danielle & Pasquale, Frank. “The Scored Society: Due Process for Automated Predictions” (2014) 89:1 Washington Law Review 1.

Chander, Anupam. “The Racist Algorithm?” (2017) 115:6 Michigan Law Review 1023.

Christian, Jon. “Why Is Google Translate Spitting Out Sinister Religious Prophecies?”

Motherboard Vice (July 20, 2018) online: https://motherboard.vice.com/en_us/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies.

Dam, Hoa Khanh et al. "Explainable Software Analytics" (2018) Presented at ICSE'18 NIER online at: https://arxiv.org/pdf/1802.00603.pdf.

Danaher, John. "The Threat of Algocracy: Reality, Resistance and Accommodation" (2016) 29:3 Philosophy and Technology 245.

Danaher, John. "Mapping the Logical Space of Algogracy", (2015) Philosophical Disquisitions online: http://philosophicaldisquisitions.blogspot.com/2015/06/how-might-algorithms-rule-our-lives.html.

Diakopoulos, Nicholas. "Algorithmic Accountability Reporting: On the Investigation of Black Boxes" (2013) Tow Center online: http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf.

Diver, Laurence & Schafer, Burkhard. "Opening the Black Box: Petri nets and Privacy by Design" (2017) 31 International Review of Law Computers and Technology 68.

Domingos, Pedro. "A Few Useful Things to Know about Machine-learning" (2012) 55:10 Communications of the ACM 78.

Doshi-Velez, Finale et al. "Accountability of AI Under the Law: The Role of Explanation" (2017) Berkman Klein Center for Internet & Society online: https://dash.harvard.edu/handle/1/34372584?show=full.

Dwork, Cynthia et al. "Fairness Through Awareness" (2012) ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 214.

Dwork, Cynthia & Roth, Aaron. "The Algorithmic Foundations of Differential Privacy" (2015) 9:3-4 Foundations and Trends in Theoretical Computer Science 211.

Edwards, Lilian & Veale, Michael. "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18.

Etzioni, Amitai. 'Is Transparency the Best Disinfectant?' (2010) 18 The Journal of Political Philosophy 389.

Fayyad, Usama. "The Digital Physics of Data Mining" (2001) 44:3 COMM ACM 62.

Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion?" (2016) online: https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

Gillespie, Tarleton. "The Relevance of Algorithms" in Tarleton Gillespie et al, eds *Media Technologies Essays on Communication, Materiality, and Society* (Massachusetts: MIT Press, 2014) 167.

Goodman, Bryce & Flaxman, Seth. "A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection" (2016) ML and the Law (NIPS Symposium 2016) online: http://www.mlandthelaw.org/papers/goodman1.pdf.

Guidotti, Riccardo et al. "A Survey Of Methods For Explaining Black Box Models" (2018) online: https://arxiv.org/pdf/1802.01933.pdf.

Harris, Sam. *Free Will* (New York: Free Press, 2012).

Health Canada, *Request for disclosure of confidential business information* (Ottawa: Health Canada, 2018) online: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/request-disclosure-confidential-business-information.html.

Hildebrandt, Mireille. "The Dawn of a Critical Transparency Right for the Profiling Era" in Jacques Bus et al, eds *Digital Enlightenment Yearbook 2012* (Amsterdam: IOS Press, 2012) 41.

Hildebrandt, Mireille. "Primitives of legal protection in the era of data-driven platforms" (2018) Submitted to Georgetown Law Technology Review for the Symposium issue on 'The Governance & Regulation of Information Platforms" online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3140594.

Hildebrandt, Mireille. "The Challenges of Ambient Law and Legal Protection in the Profiling Era" (2010) 73:3 Information Technology & Society Colloquium 428.

Introna, Lucas & Nissenbaum, Helen. "Shaping the Web: Why the Politics of Search Engines Matter" (2000) 16 The Information Society 169.

Information Commissioner's Office, Feedback request – profiling and automated decision-making, (2017) online: https://ico.org.uk/media/2013894/ico-feedback-request-profiling-and-automated-decision-making.pdf.

IJCAI-17 Workshop on Explainable AI (XAI) Proceedings. (2017) Melbourne, Australia online: http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf.

Jewell, Matthew. "Contesting the decision: living in (and living with) the smart city" (2018) International Review of Law, Computers, and Technology.

Jones, Meg Leta. "The right to a human in the loop: Political construction of computer

automation and personhood" (2017) 47:2 Social Studies of Science 216.

Just, Natascha & Latzer, Michael. "Governance by algorithms: reality construction by algorithmic selection on the Internet" (2017) 39:2 Media, Culture, and Society 238.

Karanasiou, Argyro & Pinotsis Dimitris. "A study into the layers of automated decision-making: emergent normative and legal aspects of deep learning" (2017) 31:2 International Review of Law, Computers, and Technology 170.

Kroll, Joshua et al. "Accountable Algorithms" (2017) 165 University of Pennsylvania Law Review 633.

Kroll, Joshua & Desai, Deven R. "Trust but Verify: A Guide to Algorithms and the Law" (2018) 31:1 Harvard Journal of Law and Technology 1.

Kim, Sang Ah. "Social Media Algorithms: Why You See What You See" (2017) 2 Geo Law Tech 147.

Laat, Paul de. "Algorithmic Decision-Making Based on Machine-learning from Big Data: Can Transparency Restore Accountability?" (2017) Philosophy and Technology.

Lepri, Bruno et al, "Fair, Transparent, and Accountable Algorithmic Decision-making Processes The Premise, the Proposed Solutions, and the Open Challenges" (2017) Philosophy and Technology.

Leonard, Peter. "Customer Data Analytics: Privacy Settings for 'Big Data' Business" (2014) 4:1 International Data Privacy Law.

Malgieri, Gianclaudio & Comandé, Giovanni. "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation" (2017) 7:4 International Data Privacy Law 243.

McDonald, Aleecia & Cranor, Lorrie. "The Cost of Reading Privacy Policies", (2008) 4 I/S Journal of Law & Policy for the Information Society 543 at 564.

Mendoza, Isak & Bygrave, Lee. "The Right Not to be Subject to Automated Decisions Based on Profiling" in Eleni Synodinou et al, eds *EU Internet Law: Regulation and Enforcement* (Switzerland: Springer International Publishing, 2017) 77.

Millar, Jason. "Core Privacy: A Problem for Predictive Data Mining" in Ian Ker et al. *Lessons from the Identity Trail:* Ano*nymity, Privacy and Identity in a Networked Society* (Oxford: Oxford University Press, 2009) 103.

Miller, Tim. "Explanation in Artificial Intelligence: Insights from the Social Sciences" (2017) online: https://arxiv.org/abs/1706.07269.

Mislove, Alan et al. “You Are Who You Know: Inferring User Profiles in Online Social Networks” (2010) Proceeding WSDM '10 Proceedings of the third ACM international conference on Web search and data mining 251.

Office of the Privacy Commissioner of Canada. *Consent and privacy: A discussion paper exploring potential enhancements to consent under the Personal Information Protection and Electronic Documents Act* (Report) (Ottawa: Office of the Privacy Commissioner of Canada, 2017).

O’Neil, Kathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York City, NY: Penguin Random House, 2016).

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015).

Rosen, Jeffery. “Who Do Online Advertisers Think You Are?” *New York Times* (November 30, 2012) online: https://www.nytimes.com/2012/12/02/magazine/who-do-online-advertisers-think-you-are.html.

Schnier, Bruce. *Data and Goliath* (New York: W. W. Norton & Company, 2015).

Scherer, Matthew. “Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies” (2016) 29:2 Harvard Journal of Law and Technology 354.

Selbst, Andrew & Barocas, Solon. “The Intuitive Appeal of Explainable Machines” (2018) 87:XX Forthcoming in Fordham Law Review.

Selbst, Andrew & Barocas, Solon. “Regulating Inscrutable Systems” (2017 draft paper for We Robot Conference Proceedings, University of Miami, http://www.werobot2017.com/wp-content/uploads/2017/03/Selbst-and-Barocas-Regulating-Inscrutable-Systems-1.pdf.

Selbst, Andrew & Powles, Julia. “Meaningful Information and the Right to Explanation” (2017) 7(4) International Data Privacy Law 233.

Shekhar, Amit. “What Is Feature Engineering for Machine-learning?”, *Medium* (25 November 2018) online: https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a.

Shelley, Ryan. “3 things to do after a major Google algorithm update”, *Search Engine Land* (October 18, 2016) online: https://searchengineland.com/3-things-major-google-algorithm-update-260828.

Singh, Jatinder et al. “Decision Provenance Capturing data flow for accountable systems” (2018) Computers and Society online: arXiv:1804.05741v1.

Standing Committee on Access to Information, Privacy and Ethics, Number 053, 1st Session, 42nd Parliament (Thursday, March 23, 2107) Tamir Israel at 17:16, online: https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/meeting-53/evidence;

Standing Committee on Access to Information, Privacy and Ethics, Number 054, 1st Session, 42nd Parliament (Tuesday, April 4, 2107) Ian Kerr at 16:32, online: https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/meeting-54/evidence.

Sweeney, Latanya. "Discrimination in Online Ad Delivery" (2013) 11:3 ACM Queue 1.

Temme, Merle. "Algorithms and Transparency in View of the New General Data Protection Regulation" (2017) 4 European Data Protection Law Review 473.

Thierer, Adam et al. "Artificial Intelligence and Public Policy" (2017) Mercatus Research at George Mason University, available at SSRN: https://ssrn.com/abstract=3021135.

Tufte, Edward. *The Visual Display of Quantitative Information*, 2nd ed (Connecticut: Graphics Press, 2001).

Tutt, Andrew. "An FDA for Algorithms" (2017) 69:1 Administrative Law Review 83.

Urquhart, Lachlan & Rodden, Tom. "New directions in information technology law: learning from human–computer interaction" (2017) 31:2 International Review of Law, Computers, and Technology 150.

Veale, Michael & Binns, Reuben. "Fairer Machine-learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data" (2017) 4:2 Big Data & Society 1.

Veale, Michael et al. "Some HCI Priorities for GDPR-Compliant Machine-learning" (2018) Workshop at ACM CHI'18.

Vedder, Anton & Naudts, Laurens. "Accountability for the use of algorithms in a big data environment" (2017) 31:2 International Review of Law, Computers 206.

Waldman, Ari Ezra. "Privacy as Trust: Sharing Personal Information in a Networked World" (2015) 69:3 University of Miami Law Review 559.

Wang, Hong et al. "Adversarial prediction games for multivariate losses" (2015) 2 NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems 2728.

Wachter, Sandra et al. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2016) 7:2 International Data Privacy Law 76.

Wachter, Sandra et al. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR” (2018) Forthcoming in the Harvard Journal of Law and Technology.

Weeks, Erik. “Data Mining and Kids Part One,” Wired , March 20, 2012, http://www.wired.com/geekdad/2012/03/data-mining-and-kids-part-1-thank-goodness-she-didntpay-cash/.

White House Report. “Big Data: A Report on Algorithmic Systems, Opportunity and Civil Rights” (2016) online: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Wyatt, Daniel. “The Many Dimensions of Transparency: A Literature Review” (2018) Helsinki Legal Studies Research Paper Series, No 53.

Zarsky, Tal. “The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision-making” (2016) 41:1 Science, Technology, and Human Values 118.

Zarsky, Tal. “Incompatible: The GDPR in the Age of Big Data” (2017) 47 Seton Hall Law Review 995.

Zarsky, Tal. “Transparent Predictions” (2013) 4 Illinois Law Review 1503.