© 2019, *Global Media Journal -- Canadian Edition* Volume 11, Issue 2, pp. 18-30

ISSN: 1918-5901 (English) -- ISSN: 1918-591X (Français)

Black Box Ethics:

How Algorithmic Decision-Making is Changing How We View Society and People:

Advocating for the Right for Explanation and the Right to be Forgotten in Canada

Allison Trites

Saint Paul University, Canada

Summary:

There is a growing body of evidence indicating that the algorithms informing automated decision-making systems are influenced by the fallible humans and societies who develop them. This creates a new reality in which biases are potentially just as prevalent in algorithm-made decisions as in human-made decisions. This reality has consequences for both how people see – and are seen – in the world. This article will examine some of the ethical issues surrounding automated decision-making systems and the algorithms behind them. Attention will be paid to how these algorithms affect not only the realities they purport to display but also what the data collected about and from individuals means to. On this point, the article will examine and support aspects that could be included in a framework for data ethics. Given the rapid development of automated decision-making, it is an appropriate time to assess the extent to which our reality is undermined or altered by the use of algorithmic decision-making systems in our society. To this end, in this article the Right to Explanation and Right to be Forgotten will be introduced as key policies that work to protect us from these altered realities.

Keywords: Data ethics; Privacy; Algorithms; Right to explanation; Right to be forgotten; Bias; Public sphere

There is a growing body of evidence indicating that the algorithms informing automated decision-making systems are influenced by the fallible humans and societies who develop them. The result is that a range of biases are being observed in the field of automated decision-making, where algorithms often act as mediation tools to make decisions or settle disputes: from feedback loops built into algorithms where data are collected that support existing theories, to racial discrimination — either knowingly (overt discrimination) or unknowingly (social assumptions). This creates a new reality in which biases are potentially just as prevalent in algorithm-made decisions as in human-made decisions. This reality has consequences for both how people see – and are seen – in the world.

The cloaked nature — and increasing omnipotence — of algorithms adds a further level of complexity to this new reality. This black box shields the algorithms and their resulting decisions from scrutiny and criticism while creating a new understanding of what we "know" about individuals. As a result, citizens, governments and other advocates may be less able to fully assess the algorithms, their impact on societies, and this 'new knowledge'.

This article will examine some of the ethical issues surrounding automated decision-making systems and the algorithms behind them. Attention will be paid to how these algorithms affect not only the realities they purport to display but also what the data collected about and from individuals means to. On this point, the article will examine and support aspects that could be included in a framework for data ethics. Given the rapid development of automated decision-making, it is an appropriate time to assess the extent to which our reality is undermined or altered by the use of algorithmic decision-making systems in our society. To this end, in this article the "Right to Explanation" and the "Right to be Forgotten" will be introduced as key policies that work to protect us from these altered realities.

A look at the ethical considerations of automated decision-making systems

According to Robyn Caplan and Danah Boyd's article "who controls the public sphere in an era of algorithms?"

> While the term "algorithm" has a precise technical meaning, it has entered everyday discourse as "the things computers do." In computer science, an algorithm is a step-by-step set of operations to be performed. This description makes this process seem neutral, objective, isolated, and reflective of reality. In practice, however, engineers and other company actors must make countless decisions in the design and development of algorithms. Through those decisions and relationships, subjective decisions and biases get encoded into systems.
>
> (Caplan & Boyd, 2016, p. 4)

Currently, there is not a generally accepted ethical framework for guiding those people designing automated decision-making systems or using the data these systems generate, other than what can be argued is an implied utilitarian approach in the benefits these systems may provide to society as a whole. As such, a framework with an agreed-upon set of best practices would benefit the designing and operating of these Automated Decision-Making Systems (ADMS). This is particularly so given the number and magnitude of potentially unethical uses of the systems and their related data such as the improper use of personal data, breaches of privacy and the use of data beyond what an individual had consented to. Many of these unethical uses are not simply possibilities, but have already been realized. In fact, a number of prominent scandals involving some of the world's largest and most powerful tech companies have drawn attention to the lack of ethical frameworks and regulations around data generation and use.

There are a number of hypothetical, longer-term risks to society at-large such as major shifts in the labour market. These advancements not only affect people on a day-to-day basis but may also transform the very threads that hold our modern societies together. Advancements in artificial intelligence (AI) and related systems that can process data any time, and anywhere, promote a fluid and never-ending workforce. It is estimated that more than 40% of jobs may be affected or lost in the next decades. In addition to this, more than 40% of the tasks Canadians are currently paid to do can already be automated through existing technology (Lamb, 2016, p. 5). It is estimated that the jobs most likely to be lost will be those belonging to lower-income and less-educated populations (Lamb, 2016, p. 3). Comparatively, occupations with the lowest risk of being negatively affected by automation, which are correlated with higher earnings and education, are projected to produce nearly 712,000 net new jobs between 2014 and 2024 (Lamb, 2016, p. 16). This situation is further justification for the need for an ethical framework to guide decision-making as it relates to AI. Without explicit ethical guidance, AI systems may become the most recent example of our society, through technology, helping those who may not require the assistance due to their existing access and privilege, while further disadvantaging those who are already.

In the *Cambridge Handbook of Artificial Intelligence*, Nick Bostrom, a Swedish philosopher who researches and writes on subjects like superintelligence risk, and Eliezer Yudkowsky, an AI researcher, propose the following thought experiment in their work on "The Ethics of Artificial Intelligence":

> Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?
>
> (Frankish, 2014, p. 316)

Black boxes are algorithms designed to take data inputs, process them through the decision-making course designed within them, and create the outputs, or decisions, required. The term "black box" generates powerful imagery. The information is seen going in and coming out but it is impossible, for the people from whom the information originates, to see what is done to the information while it is inside the box. The lack of transparency is further compounded by the fact that the companies that design the algorithms are often protected from sharing their algorithms publicly under proprietary laws, or 'trade secrets'. Frank Pasquale, Professor of Law at the University of Maryland Francis King Carey School of Law and author of *The Black Box Society: The Secret Algorithms That Control Money and Information*, points to the clear imbalance of the risks held by individual citizens versus those of corporations:

> The decline in personal privacy might be worthwhile if it were matched by comparable levels of transparency from corporations and government. But for the most part it is not. Credit raters, search engines, major banks, and the TSA (Transportation Safety Authority) take in data about us and convert it into scores, rankings, risk calculations, and watch lists with vitally important consequences. But the proprietary algorithms by which they do so are immune from scrutiny, except on the rare occasions when a whistleblower litigates or leaks.
>
> (Pasquale, 2016, p. 4)

Initially, AI was packaged to society as an unemotional and, therefore impartial, judge, touted as a great equalizer of sorts. All humans could be equal in the eyes of an automated system, one without the ability to see skin colour or other factors that might traditionally have biased human decision-making. Judgements would be based only on the facts (data) and decisions (outputs) could therefore be impartial. One of the reasons for this perception is that these systems are based on empirical evidence, which is generally perceived to be impartial and objective. However, those who study the philosophy of science debate whether the scientific method truly is "value-free". According to John Cheney-Lippold, explaining the relationship between digital media, identity, and the concept of privacy, "data does not naturally appear in the wild. Rather, it is collected by humans, manipulated by researchers, and ultimately massaged by theoreticians to explain a phenomenon. Who speaks for data, then, wields the extraordinary power to frame how we come to understand ourselves and our place in the world" (2017, p. 107).

Recent studies examining automated decision-making systems, as introduced in the next section of this article, have shown that embedded biases and values of those conducting the analyses have direct effects on the decisions reached by these systems. We are building these machines assuming their impartiality, but are coding partial and biased values into them.

Ethical considerations: Examples

As will be demonstrated through the following examples of black box algorithms and their effects on the individuals subjected to them, the data inputs are collected, analysed and leveraged to make determinations or judgements. The judgements have direct effects on not only how the individuals are perceived but also how, in turn, their rights and future potentials are affected. As simply put by Pasquale, "Pattern recognition is the name of the game— connecting the dots of past behaviour to predict the future" (Pasquale, 2016, p. 20).

Evaluating Automated Decision Making Systems – COMPAS

Several studies have been conducted examining and reviewing the results of automated decision-making systems and their effects. One example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, a program available in the US created by a commercial vendor that uses machine learning to address biases in the judicial system. The program replaces a judge's decision in sentencing in order to remove the potential for personal bias. The underlying premise is that an algorithm can help to mitigate the 'unchecked power' held by judges. "Judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood of becoming a recidivist" (Larson, Surya, Kirchner & Angwin, 2016). ProPublica, an American non-profit organization based in New York City that "describes itself as a non-profit newsroom that produces investigative journalism in the public interest" (propublica.org) conducted an evaluation of the outputs and resulting sentences that the COMPAS program recommended. "The study found that shifting the sentencing responsibility to a computer does not necessarily eliminate bias; it delegates it and often compounds it" (Larson, Surya, Kirchner & Angwin, 2016).

Overall, the analysis concluded that the program's validity had not been robustly tested prior to implementation. The validity was assessed in "about 1 or 2 studies and often by the same people who developed the instrument" (*Ibid*). It is important to know that "no one knows how COMPAS works; its manufacturers refuse to disclose the proprietary algorithm. We only know the final risk assessment score it spits out which judges may consider at sentencing" (*Ibid*).

The issue here is not only the use of the algorithm, but also the abdication of responsibility by judges in rendering sentences. While the human bias of judges is a flaw in the current judicial system, it can be argued that at least these biases are more transparent than those contained within black box algorithms. According to Ellora Israni, J.D. candidate at Harvard Law School and former software engineer at Facebook, "This is precisely why states are abdicating the responsibility for sentencing to a computer. Use of a computerized risk assessment tool somewhere in the criminal justice process is widespread across the United States, and some states, such as Colorado, even require it. States trust that even if they cannot themselves unpack proprietary algorithms, computers will be less biased than even than [humans]. But shifting the sentencing responsibility to a computer does not necessarily eliminate bias; it delegates and often compounds it" (Israni, 2017). The use of statistical analysis as the only factor in making a decision is problematic in that it takes out the weighting of socially relevant factors that would otherwise be considered by a human. "Algorithms also lack the human ability to individualize. A computer cannot look a defendant in the eye, account for a troubled childhood or disability, and recommend a rehabilitative sentence. This is precisely the argument against mandatory minimum sentences — they rob judges of the discretion to deliver individualized justice — and it is equally cogent against machine sentencing" (Israni, 2017).

Algorithms like those found in COMPAS, increase the exposure to vulnerabilities of individuals subjected to them. According to Pasquale "the problem of collateral consequences is well known in the criminal justice system. Once someone has been convicted of a crime (or pleaded guilty), that stigma will often preclude him from many opportunities—a job, housing, public assistance, and so on—long after he has 'paid his debt to society'" (Pasquale, 2016, p. 41). This stigmatization is problematic for those with already increased vulnerabilities, as will be discussed later in this article with the focus on vulnerabilities.

Evaluating Automated Decision Making Systems: PREDPOL

PredPol is another example of an automated decision-making system that is currently in use in the United States. The program collects "historical crime data" and generates predictions about where crimes are more likely to occur. Police departments then use this data to ensure that these neighbourhoods are more frequently patrolled. A pilot study of the PredPol system, in Santa Cruz, California, resulted in a claim that burglaries went down by 23% (O'Neill, 2017, p. 84), as described by Cathy O'Neill, an American mathematician who writes on data science and the effects of predictive algorithms on our society. "Predictive programs like PredPol are all the rage in budget-strapped police departments across the country. Departments from Atlanta to Los Angeles are deploying cops in the shifting squares and reporting falling crime rates […] Like those in the rest of the Big Data industry, the developers of crime prediction software are hurrying to incorporate any information that can boost the accuracy of their models" (O'Neill, 2017, p. 85).

While reduced crime rates are a positive result being attributed to a program like PredPol, the criticism the program also receives centres around the fact that increased police patrols have also had negative effects on these communities. While burglaries and thefts are more nefarious crimes that were caught by the program, the increased police presence also resulted in arrests and police intervention in lesser crimes, or "nuisance crimes" such as vagrancy, and sometimes labelled by police as 'antisocial behaviour', "which would go unrecorded if a cop weren't there to see them"

(O'Neill, 2017, p. 86). This antisocial behaviour was seen as a neighbourhood destructor of sorts, creating an atmosphere of disorder and as a result, "scaring all the law-abiding citizens away" (O'Neill, 2017, p. 87).The problem is that these crimes, labelled as 'nuisance crimes' are more likely to occur in impoverished neighbourhoods. Further, without focusing on more serious crimes, these nuisance crimes work to skew the data and generate a positive-feedback loop within the PredPol system which leads to further increases in police presence in these communities which in turn results in the potential for further arrests therein. The analysis of risk, in this situation, is a primarily subjective judgement coded into algorithms – yet appears and is marketed as an objective decision without bias or prejudice. "Nevertheless, to have someone knock on your door because your data is seen to be 'at risk' reaffirms some of the worst fears we might have about this new, datafield world: our data (Cheney-Lippold, 2017: 488).

This feedback loop reaffirms the 'problem areas' in a city and reconfirms the negative connotations of a neighbourhood. In this way, a paternalistic relationship between police and impoverished or marginalized neighbourhoods and individuals is reinforced[, as is the social position of the impoverished or marginalized neighbourhood and individual. "The policing itself spawns new data, which justifies more policing. And our prisons fill up with hundreds of thousands of people found guilty of victimless crimes. Most of them come from impoverished neighborhoods, and most are black or Hispanic. So even if a model is color blind, the result of it is anything but. In our largely segregated [US] cities, geography is a highly effective proxy for race" (O'Neill, 2017, pp. 86-87).

Amazon Recruitment

In 2014, Amazon created a recruitment analysis tool that, deploying machine learning, would analyze applicant's resumes and search for "top talent", providing each with a rating from 1 to 5 stars. After tests and trials, the company realized that the results skewed towards male candidates, as there were more resumes received from male candidates (due to a higher percentage of men in the tech industry) in the system for the program to review and learn from. "In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's", as in "women's chess club captain." And it downgraded graduates of two all-women's colleges" (Huang, 2017).

While Amazon attempted to rewrite the program in order for it to be able to review and rank candidates in a gender-neutral way, it was eventually disbanded. The reports concluded that there was no way to ensure that discriminatory assumptions or results would be avoided or that "the machines would not devise other ways of sorting candidates that could prove discriminatory" (Huang, 2017). According to an article from Reuters on the Amazon recruitment tool:

> The company's experiment […] offers a case study in the limitations of machine learning. It also serves as a lesson to the growing list of large companies including Hilton Worldwide Holdings Inc and Goldman Sachs Group Inc that are looking to automate portions of the hiring process. Some 55 percent of U.S. human resources managers said artificial intelligence, or AI, would be a regular part of their work within the next five years, according to a 2017 survey by talent software firm CareerBuilder. Employers have long dreamed of harnessing technology to widen the hiring net and reduce reliance on subjective opinions of human recruiters. But computer scientists such as Nihar Shah, who teaches machine learning at Carnegie Mellon University, say there is still much work to do. "How to ensure that the

> algorithm is fair, how to make sure the algorithm is really interpretable and explainable - that's still quite far off," he said.
>
> (Huang, 2017)

The danger with black box algorithms is clear: negative decisions and outcomes can result from embedded biases. This danger is further exacerbated by the belief that these outcomes are not the results of discriminatory practices but are instead facts supported by 'evidence'. As clearly portrayed by Pasquale, "Bad inferences are a larger problem than bad data because companies can represent them as "opinion" rather than fact" (Pasquale, 2016, p. 32). While there are many positive effects that these examples can point to in order to support their necessity (from reducing crime rates to reducing discrimination in reviewing criminal cases or resumes), the fact remains that these positive effects are undone by the embedded discriminations in the algorithms themselves and the ripple effects they cause for those individuals affected. While the reduction of crime may be of critical importance in a community or society, the reality is that in certain instances this is done on the backs of marginalized, vulnerable or already discriminated-against persons. As stated by Israni, "Machine learning algorithms often work on a feedback loop. If they are not constantly retrained, they "lean in" to the assumed correctness of their initial determinations, drifting away from both reality and fairness. As a former Silicon Valley software engineer, I saw this time and again: Google's image classification algorithms mistakenly labeling black people as gorillas, or Microsoft's Twitter bot immediately becoming a "racist jerk'" (Israni, 2017).

Who we (think) we are and what we (think) we know

In their article, "Who Controls the Public Sphere in an Era of Algorithms?", Caplan and Boyd, address the issues surrounding the role social media plays in the shaping of the realities of its participants. In recognizing the benefits that come from having the reach that the algorithms have through social media platforms (e.g., increased voter turnout related to 'nudging'), the article also points to the way that these algorithms are introducing new variables into how and why people gather the information they access to form their opinions. Effectively, the automated decision-making systems supported by the algorithms are affecting the interpersonal decisions we make as humans, supported by the information available to us in our own spheres. "Framing, journalistic bias, and variables affecting the 'newsworthiness' of an event or set of information in any given time or place, also affect what and how information is filtered into public view. As media have become networked, researchers have pointed to additional set of processes and mechanisms that are shaping public life and the production, dissemination, and consumption of news and information – namely algorithms, data, and automation" (Caplan & Boyd, 2016, p. 2). This points to how data and algorithmic design shape what information and news we see and when we see it.

The authors discuss to Jürgen Habermas's works on the public sphere, which according to him is, "the 'realm of our social life in which something approaching public opinion can be formed,' and is the space which mediates between "society and state, in which the public organizes itself as the bearer of public opinion"" (Caplan & Boyd, 2016, p. 2). The public sphere is arguably a place where public opinion is enabled through access to information and larger-scaled public discourse. Habermas asserts that, while early on mass media became an arm of government and corporate interests, there was a need for an "ideal form of public sphere" (Caplan & Boyd, 2016, p. 3). wherein media and public discourse could and would play a role in shaping public opinion in a positive way, participating in democracy. It was argued that Habermas's theory, has been dismantled due to a central flaw in his posited ideal: "Even in its ideal form, this public sphere would have been exclusionary and homogenous […]" (Caplan & Boyd, 2016, p. 3). due to the power dynamics that exist in and across societies. The decisions made as to what would be included in public discourse, either by the media or cultural values, by a society would not be free of the intervention of those holding the power in that society.

Similarly, when algorithms are engaged to collect and enable data as decision-making, as seen in the examples earlier in this article, there are human decisions being made behind the automated decisions. As pointed to by Caplan and Boyd for example, "most recently, a number of scholars and journalists have focused on the shaping of the public sphere through more covert means, beyond our awareness, through the design and iteration of algorithms and automated media" (Caplan & Boyd, 2016, p. 3). The reasoning behind these algorithms is to often simplify a process or to condense fact into a resulting reality. However, according to Caplan and Boyd, this reality is subjected to the earlier made decisions by the algorithm's designers or the companies they represent, about what information/data is available to them and then even more so, what parts of that they chose to include and how the resulting decisions are ultimately portrayed. The results can, and as has been shown, be rife with encoded biases in the decision-making systems due to the fact that all humans have a level of bias built into their own decision-making systems. All of this is often done behind a veiled curtain of secrecy and proprietary rights.

While Caplan and Boyd's article focusses on social media algorithms and their influence on public discourse, it is useful to consider this in a bigger picture context over data use, collection and storage overall. This data is often collected through an oft-biased system (as seen in the examples given earlier) for use in a biased algorithm and resulting in biased decision-making. Returning to Habermas, Caplan and Boyd point to the 'new public sphere':

> Critics of Habermas have challenged the bourgeois public sphere that he articulates, arguing that he imagined an ideal that never was. In many ways, the same can be said of those seeking an Internet enabled public sphere. Technology has reconfigured aspects of the public sphere, but perhaps not always in the way that many would like. Yet, it is important to consider who is included in this new configuration, who is not, and how this is like or unlike previous instantiations.
>
> (Caplan & Boyd, 2016, p. 8)

…And What Can Be Done About It

While Caplan and Boyd focus on social media and the use of a media based in technology and backed by algorithms in the enabling of a 'new public sphere', there is great overlap that can be pulled into the discussion around the use and collection of data to create a 'new public reality'. All humans operate with a level of bias and these biases are embedded in inaccessible algorithms that spread and further entrench biases, resulting in greater inequalities and vulnerabilities for many people. Government regulation and other forms of systemic oversight are required to mitigate these risks.

The GDPR (General Data Protection Regulation) is an EU legislation that imposes on data controllers requirements around the data collected from subjects and how it is used. The GDPR defines personal data as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (GDPR). There is an onus, created through the legislation, on the data collector to act in a more transparent way to ensure the control over an individual's data rests with the individual.

Among the many ethical considerations surrounding privacy and an individual's data use and storage, the Right to Explanation and the Right to be Forgotten are being debated – and contested – in the EU, which is currently examining rights surrounding its citizens' data in the aforementioned GDPR. In reaction to this opaqueness, the Right to Explanation (RtE) and the Right to be Forgotten (RtbF) are key policy responses being debated and discussed by governments to address these risks associated with automated decision-making systems. The RtE is critical in that it walks the line between an individual's right to understand why a decision was made for or against them and a company's proprietary rights. Similarly, the RtbF grants a level of ethical consideration for an individual's privacy and control over what aspects of themselves they want visible to the world in an age where internet-driven memories last forever. These Rights underscore the importance of ethical challenges in the use of algorithms and data for automated decision-making systems. They are important to examine as the discourse around them become the intersection between the importance of data collection and the need to assist in the determination of where a person's privacy ends and where (or whether) the benefits that data outweigh the effects of the invasion used to extract it. Debate on the legislation brings up questions about whether it implies a Right to Explanation (Right to Explanation) for these subjects, where an individual would have access to "human intervention", the ability to contest the situation and the right to "obtain an explanation of the decision reached after such assessment and to challenge the decision" (Intersoft Consulting, n.d.). These are the considerations currently being debated in the EU through their development of policies on data protections, such as the GDPR.

The question central to the debate around the Right to be Forgotten (Right to be Forgotten) and the GDPR is whether current and new regulations around an individual's right to access and protect their data then creates a Right to Explanation for any decisions made through automated decision-making systems for that individual. To paint the picture, Frank Pasquale demonstrates the problem with the collection, storage and sharing of personal data with the advancements in ADMSs: "Profiling may begin with the original collectors of the information, but it can be elaborated by numerous data brokers, including credit bureaus, analytics firms, catalogue co-ops, direct marketers, list brokers, affiliates, and others. Brokers combine, swap, and recombine the data they acquire into new profiles, which they can then sell back to the original collectors or to other firms. It's a complicated picture, and even experts have a tough time keeping on top of exactly how data flows in the new economy" (Pasquale, 2016, p. 32).

Given the rapid development of automated decision-making, it is an appropriate time to assess the extent to which these rights should be introduced into Canadian policies and legislations. Search engines and websites are notorious collectors and users of individual data. From a Canadian perspective, the approaches to data protection from Canada's Privacy Commissioner— those of de-indexing, source takedown and privacy education — are a step in the right direction. However, as mentioned, they fall short of the intention of a true Right to Explanation and Right to be Forgotten.

As outlined in the Office of the Privacy Commissioner's Annual Report to Parliament, "Trust but Verify", the focus of the House of Commons Committee is to maintain Canada's adequacy status with the EU. The Report in fact references a recent Canadian-based court case, Google Inc. v. Equustek Inc. (2017 SCC 34), the connection with the Right to be Forgotten debate being that the court case led to the precedent that a search engine can be ordered to "remove (de-index) information from its search results" (Office of the Privacy Commissioner of Canada, 2018, p. 14). The impact this case has, according to the report, is on the direction that privacy legislation has and may further take towards supporting a Right to be Forgotten in the future. However, the Report stops short of calling for the Rights and their benefits to individuals in maintaining autonomy over their own data. As per the previously reviewed Guidelines and Draft Position Paper on Online Reputation, Canada is holding firm, so far, at actions such as de-indexing as the available opportunities for individuals to maintain provenance over their data. The GDPR, as discussed, goes much further in explicit Rights to Explanation and to be Forgotten.

Indeed, Canada should aim much higher than being adequate in its provisions for its citizens in maintaining privacy and control over their data in the face of black box algorithms**.** I believe Canada should and needs to establish and enforce a true Right to Explanation and Right to be Forgotten if it hopes to address the ethical considerations for data collection and protect its citizens, as referenced in the Draft Paper on Online Reputation. Canada and the PIPEDA need to move in the direction of strengthened legislation, like that of the GDPR, to ensure the autonomy of all of its citizens.

Bringing it all together: how we can keep the focus on people as people

As has been discussed in this article, the use of algorithms to make decisions is changing not only how people are seen by the world but also how people see the world. There is much that can and needs to be done in this new frontier of ethical investigation and research. The biases within these algorithms are not new to our societies, however their embedded nature make them harder to reveal and to control for. This makes the sharing, collecting and applying of knowledge and facts that are not skewed or tainted by these biases, both our own and that of others around us, even more difficult today than it has been previously.

As my analysis demonstrates, without an applied ethical framework, the current model of black box algorithms and their effects on society are, as a result, deleterious for the individuals subjected to them. We, as a society, cannot permit them to be further propagated without creating and enacting a framework through which their evolution must be guided. To do so would be detrimental to not only those with vulnerabilities in our society but, left, unchecked, will work to destroy our society's already diminished capacity for privacy and autonomy. Two of the steps towards a framework can be found in the Right to Explanation and the Right to be Forgotten, as outlined in the GDPR. These Rights would allow for technologies and ADMSs to advance as they currently are but through a system of checks and balances, requiring the companies and governments deploying them to be held accountable by society and responsible to the individuals upon whose data companies develop programs and therefore generate profit. These technologies, through these new levels of transparency, would therefore be able to declare, and rightfully so, the benefits they claim to provide to society through this new required and maintained level of scrutiny.

Data collection does not have to be the proverbial boogeyman, where only bad things can result from it. With protections like the Right to Explanation and the Right to be Forgotten, which as demonstrated can work towards ensuring that our identities, realities and other ethical considerations are accounted for during data collection, storage and use processes, data collection can be positive building blocks for policies and programs or for the algorithms themselves. It is the ethical considerations and removal of the blank cheque given to these programs that is at issue.

There is a propensity in our society to always be looking for the new advancement that will enhance our lives, our jobs, and our abilities. The hope is to expand our knowledge and our influence into previously inaccessible areas of the globe and at a previously unattainable speed. These enhancements also allow for quicker decisions and actions, supporting businesses and governments in their efforts to manage growing populations and problems. However the propensity for these enabling tools also works to relegate all of us to pieces of data to be faster and more easily consumed, judged and controlled.

This article argues that there are many ethical concerns to be considered around the use of algorithms and how they introduce covert biases that affect individual lives on many levels. In addition to the examples provided in the beginning of this article, such as those of COMPAS or PredPol these algorithms more specifically can also affect our interpretation of own understanding of the world around us and our own realities through our use of social media, as introduced through introducing Caplan and Boyd's critique of Habermas. Through this critique, this article demonstrated that it is critically important that oversight be put into the protection of citizens from algorithms reshaping their realities, both in overt and covert ways. This oversight can be found in the policies of the Right to Explanation and the Right to be Forgotten, as important tools to address these ethical concerns and as a way forward.

References

Bartolini, C. & Siry. L. (2016). "The Right to be Forgotten in the Light of the Consent of the Data Subject." *Computer Law & Security Review: The International Journal of Technology Law and Practice*, *32*(2), 218–237.

Bostrom, N., & E. Yudkowsky. (2014). "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Frankish, K., & Ramsey, W.M. (pp.316—334), Cambridge: Cambridge University Press. Available online at: https://doi.org/10.1017/CBO9781139046855. [Accessed 27 April 2019].

Office of the Privacy Commissioner of Canada. (2017) "Trust But Verify: Rebuilding trust in the digital economy through effective, independent oversight". *2017-18 Annual Report to Parliament on the Personal Information Protection and Electronic Documents Act and the Privacy Act.* Available online at: https://www.priv.gc.ca/en/opc-actions-and-decisions/ar_index/201718/ar_201718/. [Accessed 27 April 2019].

Caplan, R., & Boyd, D. (2016, May 13). "Who Controls The Public Sphere In An Era Of Algorithms?" *Data Society.* Available online at: https://datasociety.net/pubs/ap/MediationAutomationPower_2016.pdf. [Accessed 27 April 2019].

Cheney-Lippold, J. (2017). *We Are Data: Algorithms and The Making of Our Digital Selves*. New York: New York University Press.

Daigle, T. (2019, 26 Sept). "Europeans have a 'right to be forgotten' online. Should Canadians?". *CBC.* Available online at: https://www.cbc.ca/news/technology/right-to-be-forgotten-canada-eu-court-1.5297528. [Accessed 26 Sept. 2019].

Université de Montréal. (2017). "Declaration of Montréal for a Responsible Development of AI."

Declaration of Montreal; Al for a Responsible Development of AI. Available online at: https://www.montrealdeclaration-responsibleai.com. [Accessed 27 April 2019].

Geist, M. (2018). "Does Canadian Privacy Law Apply to Google Search?". *Michael Geist.* Available online at: http://www.michaelgeist.ca/2018/10/does-canadian-privacy-law-apply-to-google-search/. [Accessed 27 April 2019].

Huang, H. (2017). "Dominated by Men." *Reuters Graphics.* Available online at: https://fingfx.thomsonreuters.com/gfx/rngs/AMAZON.COM-JOBS-AUTOMATION/010080Q91F6/index.html. [Accessed 24 Feb. 2019].

BBC. (2018, 13 Apr.). "Google Loses 'Right to Be Forgotten' Case." *CBC.* Available online at: https://www.bbc.com/news/technology-43752344. [Accessed 27 April 2019].

Granville, K. (2018). "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens." *The New York Times*. Available online at: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html. [Accessed 27 April 2019].

Israni, E.T. (2017). "When an Algorithm Helps Send You to Prison." *The New York Times*. Available online at: https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html. [Accessed 27 April 2020].

Kaminski, M. E. (2018) "The Right to Explanation, Explained." *University of Colorado Law Legal Studies Research Paper, 18-24.* Available online at: https://ssrn.com/abstract=3196985. [Accessed 27 Apr. 2019].

Lamb, C. (2016) "The Talented Mr. Robot: The impact of automation on Canada's workforce." *Brookfield Institute*. Available online at: http://brookfieldinstitute.ca/research-analysis/automation/. [Accessed 27 April 2019].

Larson, J., Surya, M., Kirchner, L., & Angwin, J. (2016). "How We Analyzed the COMPAS Recidivism Algorithm" *ProPublica*. Available online at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed 27 April 2019].

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown/Archetype.

Google Inc. (n.d.) "Our Principles." *Google AI.* Available online at: https://ai.google/principles/. [Accessed 27 Apr. 2019].

Pasquale, F. (2016). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Intersoft Consulting. (n.d.). "Recital 71 - Profiling." *General Data Protection Regulation (GDPR).* Available online at: https://gdpr-info.eu/recitals/no-71/ [Accessed 24 Feb. 2019].

Thomson Reuters (24 Sept. 2019). "Google Wins 'Right to be Forgotten' fight with France." *CBC.* Available online at: https://www.cbc.ca/news/technology/google-wins-right-to-be-forgotten-fight-with-france-1.5294957. [Accessed 24 Sept. 2019].

State v. Loomis., 130 Harv. L. Rev. 1530 (2019). Available online at: https://harvardlawreview.org/2017/03/state-v-loomis/. [Accessed 27 Apr. 2019].

Winston, M.E., & Edelbach, R. (2009). *Society, Ethics, and Technology*. 4th ed.. Wadsworth: Cengage Learning.

About the Author

Allison Trites is a recent graduate of Saint Paul University in Ottawa, Ontario, with a Master's in Public Ethics. Her research interests are in data ethics, particularly in the ethical considerations around both public and personal data collection, use and storage, with special emphasis on those populations with increased vulnerabilities. She previously graduated from Carleton University in Ottawa, Ontario with a B.A. in Psychology.

Citing this article:

Trites, Allison. (2019). Black Box Ethics: How Algorithmic Decision-Making is Changing How We View Society and People. Advocating for the Right for Explanation and the Right to be Forgotten in Canada, *Global Media Journal -- Canadian Edition*, *11* (2), pp. 18-30.