Saint Paul University

BLACK BOX ETHICS:

WHY THE RIGHTS TO EXPLANATION AND TO BE FORGOTTEN ARE ETHICALLY CRITICAL COMPONENTS FOR VULNERABLE POPULATIONS

Allison Trites

EPE 6999

Dr. Monique Lanoix

March 1, 2019

TABLE OF CONTENTS

BLACK BOX ETHICS: WHY THE RIGHTS OF EXPLANATION AND TO BE FORGOTTEN ARE ETHICALLY CRITICAL COMPONENTS FOR VULNERABLE POPULATIONS

Automated decision-making systems are becoming more and more prevalent in our society. These systems are often purported by their developers to remove human bias from decision-making in a variety of sectors from financial systems (e.g., credit card and mortgage applications) to Human Resources (e.g., screening resumes and matching competencies for employers). While the assertion of bias reduction is a commonly cited benefits of these automated decision-making systems, recent research and scandals in the news demonstrate that it is not safe to assume that these systems are free from bias.

There is a growing body of evidence indicating that the algorithms informing automated decision-making systems are influenced by the fallible humans who develop them. The result is that a range of biases is being observed in the field of automated decision-making: from feedback loops built into algorithms where data are collected that support existing theories, to racial discrimination — 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. The risks to citizens, however, may be heightened due to the incorrect assumption that these algorithms are unbiased.

The cloaked nature — and increasing omnipotence of — algorithms adds a further level of complexity to this situation. Many algorithms are developed by private companies and individuals who which cite proprietary rights — or ‘trade secrets’ — as protections and justifications for confidentiality around their algorithms. These rights enable developers to keep the rationales used for making decisions, built into the algorithms, in a ‘black box’. This black

box shields the algorithms and their resulting decisions away from scrutiny and criticism. As a result, citizens, governments and other advocates are unable to fully assess the algorithms and their impact on societies.

Margot Kaminski from the University of Colorado and Yale Law Schools, introduced later in this paper, lays out the global issues surrounding black box algorithms. “Scholars and civil society groups on both sides of the Atlantic have been calling for algorithmic accountability: laws governing decision-making by complex algorithms, or AI. Algorithms can be used to make, or greatly affect, decisions about credit, employment, and more. Algorithmic decision-making can be opaque, complex, and subject to error, bias, and discrimination, in addition to triggering dignitary concerns” (2). 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. 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. These are the considerations currently being debated in the EU through their development of policies on data protections, such as the General Data Protection Regulation.

This thesis will examine the ethical issues surrounding automated decision-making systems and the algorithms behind them. In this paper, I will apply to this new field of data ethics a framework by feminist bioethicists who argue for policies that protect the autonomy of

'vulnerable populations'. Utilizing a defined understanding of vulnerability – and its relationship with autonomy – will provide a clearer understanding of the role of consent for vulnerable populations. The current model of consent often applied in algorithmic decision-making is flawed, in that it makes assumptions that undermine autonomy, works to effectively remove the ability for consent for vulnerable populations, and, in fact, perpetuates biases, socialized oppression and unwarranted paternalism. An ethical framework that focuses on vulnerable populations and socialized oppressions that is then applied to data collection – in the form of the RtE and the RtbF – will work to facilitate and support autonomy for all people. This is particularly true for those individuals with increased numbers of vulnerabilities in our society and requires that the autonomy, through consent or the ability to object, is enabled. It is for the fact that the issue of consent is so problematic that the RtE and the RtbF are critical elements for any individual, vulnerable or not, in achieving autonomy.

A feminist bioethics framework, as proposed by Wendy Rogers, *et al.* in, "Why Bioethics needs a Concept of Vulnerability", and Sisti and Stramondo in, "Competence, Voluntariness, and Oppressive Socialization: A Feminist Critique of the Threshold Elements of Informed Consent", is particularly appropriate for analysing these issues as these frameworks centre on vulnerability and autonomy. Feminist theorists in bioethics argue that "oppressive socialization undermines the liberal model of autonomy" (Donchin). They provide to the larger field of bioethics a focus on marginalized groups that are at risk of being overlooked in favour of a larger, less diverse group. Throughout the development of bioethical theory, "during the 1980s, feminists […] argued that bioethics was developing in a way that gave too little attention to gender-specific disparities in health care research and therapy, or to the effects of other power disparities, such as class and ethnicity, on quality of health care" (Donchin). By the 1990's, feminist bioethics rose

in distinction through its focus on women's health and biology among other topics. A feminist bioethical framework is appropriate for examining how automated decision-making systems affect an individual's autonomy as it centres on the vulnerabilities of an individual while not holding the individual as the sum of their vulnerabilities. In perceiving an individual as a whole person beyond these factors of disadvantage, feminist bioethical theorists would argue, paternalistic and oppressive relationships can be dismantled and the autonomy of the individual is better supported.

> Notably, feminist discourse highlights the way in which hierarchical rankings that categorize people by sex, race, ethnicity, age, disability, or susceptibility to genetic disease, can perpetuate unjust practices in health and social care, research, and public health. Some feminists integrate cross-disciplinary analysis of structural and social frameworks that divide and marginalize people with insights from the women's health movement, others concentrate their analysis on a specific axis of oppressive practice, but all recognize interrelationships among such practices (Donchin).

With this information and the proposed feminist bioethical frameworks, I will conduct a case study on a program currently being implemented in Ontario, the Brief Mental Health Screener (BHMS) by HealthIM, as an example of an emerging system designed to be used with persons in a mental health crisis. I will examine and question what the program will mean to vulnerable populations in Canada — those that will be most greatly affected, both positively and negatively, by the outcomes prescribed by the program. The case study will, in the form of a small vignette, consider how interactions between police officers and the people with vulnerabilities in our society, specifically mental health issues, are being exposed to algorithms.

As will be demonstrated through the case study, algorithms are used to assist and inform a police officer how to engage the individual, assess the situation, determine the correct course of action, and inform other front-line service workers (e.g., hospitals) of the incident. The case study will be used to demonstrate the ethical considerations of vulnerability, autonomy and informed consent that need to be measured when collecting, using and sharing the data of vulnerable populations between sectors. Applying the proposed framework, I will examine whether the RtE and the RtbF are, or need to be, critical pieces of privacy and data ethics legislation as they foster autonomy and therefore actual consent — in particular for vulnerable populations. In the new age of algorithmic decision-making, those parties promoting these systems continue to purport that they remove bias and provide equality in access and decision-making for society as a whole. In this paper, I will argue that this is a false, and dangerous, conclusion.

This thesis will do the following: the situation around and important concepts in Artificial Intelligence (AI) and Automated Decision-Making Systems (ADMS) will be introduced and defined, focusing on Black box algorithms and the problems surrounding these systems. From here, the case study will be introduced, introducing the reader to specific aspects of the concerns around black box algorithms, as an example to be kept in mind for the remainder of the thesis, culminating in the subsequent analysis at the end of the paper. Following this, a feminist bioethical framework on vulnerability and consent will be outlined and reviewed. After the introduction of the case study and proposed framework, current Canadian legislation surrounding data collection and privacy will be reviewed, specifically the Personal Information Protection and Electronic Documents Act (PIPEDA) and a working paper from the Office of the Privacy Commissioner of Canada, providing the reader a Canadian-based context. Next, the concepts of the RtE and then the RtbF will be introduced and discussed in relation to the collection and

storage of personal information. At this point, an analysis will be conducted, bridging the RtE and the RtbF with the feminist bioethical framework on vulnerability and consent and culminating in a discussion on and a revisiting of the case study. Using the case study, the paper will conclude how the RtE and the RtbF work to best support vulnerable populations and their ability to achieve autonomy and how the BMHS application could be adapted to best accommodate this objective.

1. ARTIFICAL INTELLIGENCE (AI) AND AUTOMATED DECISION-MAKING SYSTEMS (ADMSs)

a. Introduction and review of concepts

I will begin with an introduction of the concepts and definitions involved when discussing Artificial Intelligence (AI) and related systems. AI itself has become a misnomer of sorts, in that it has become a catch-all phrase. This in itself can be misleading and dangerous because it allows for a glazing over of potential risks to our societies' freedoms and securities.

There are some general definitions of AI that are often used, demonstrating its ambiguous presence in our society today. One of these is, "artificial intelligence is a computerised system that exhibits behaviour that is commonly thought of as requiring intelligence" (Van Duin). Another is "artificial Intelligence is the science of making machines do things that would require intelligence if done by man" (Van Duin). In these definitions, the concept of intelligence refers to some level of ability to plan, reason and learn, sense and build some kind of perception of knowledge and communicate in natural language. The founding father of AI, Alan Turing, put forth: "AI is the science and engineering of making intelligent machines, especially intelligent computer programs" (Van Duin). This was in the 1950s. We can see from then to now, the concept is still very high-level and nondescript, leading to a lack of understanding in the general public about the scope and realities of AI and underscoring societies' seemingly inability, or lack of interest, in comprehending and anticipating the risks associated with it and related systems.

There are some specific concepts related to AI systems that are important to understand as background to this thesis. Narrow Artificial Intelligence (NAI) refers to almost all of the current AI applications. It is narrow because it is designed to solve or assist in a specific

problem, and this is all that it is capable of doing. As an example, NAI refers to a specific algorithm being designed for a specific task, as will be seen in Machine Learning, and while it will be remarkably better and/or faster at that task than a human, it doesn't have the capabilities of humans do to perform other tasks. General Artificial Intelligence (GAI) is described as the "holy grail" (Van Duin) of AI, "a single system that can learn and then solve any problem it encounters. This is exactly what humans do: we can specialise in a specific topic, from abstract maths to psychology and from sports to art. We can become experts at all of them" (Van Duin).

As a form of NAI, machine learning is the process where, through the algorithm written for it, the system is trained and therefore learns and processes certain information. It can learn to identify patterns and can even improve itself by creating feedback loops which reinforces what it learns and how it learns. In machine learning, the algorithm will never internalize or postulate further on what it was trained to do, as a human brain would. Automated decision-making systems (ADMSs), the focus of this thesis, fall into the category of NAI.

Other forms of AI are: Cognitive analytics, robotics and Smart Machines. Cognitive analytics is a system capable of extracting information from unstructured data by 'mining' concepts and relationships into a knowledge base; Robotics, which uses AI to bring it from being a simple machine to a machine that can learn how to do more than just a simple task, like that seen in self-driving cars. Smart Machines, are further along the spectrum of AI, where the main aspect of its differentiation is in its autonomy.

> Smart Machines are systems that – to some extent – are able to make decisions by themselves, requiring no human input. Cognitive Analytics systems as well as robots, or any kind of AI, can be called Smart Machines, as long they adhere to this rule (Van Duin).

It is important to recognize this spectrum of technological advancement, and where these systems may evolve to, when considering the ethical implications involved. There is value is addressing a these advancements early in their evolution, when potential problems may be more manageable. It is for this reason that an ethical framework for ADMSs, at this early stage in their development, is so important.

b. Ethical Considerations

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[1], 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 ADMSs. 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 are becoming realities, as will be discussed in Chapter 2. 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[2]. Jurisdictions like

[1] Recent efforts, such as the "Montreal Declaration for a Responsible Development of Artificial Intelligence" ("Declaration of Montréal for a Responsible Development of AI." *Declaration of Montreal; AI for a Responsible Development of AI*, https://www.montrealdeclaration-responsibleai.com. Accessed 27 Apr. 2019.) and "Google AI" ("Our Principles." *Google AI*, https://ai.google/principles/. Accessed 27 Apr. 2019.) are taking steps towards the recognition of the need for social responsibility around developments in Artificial Intelligence. These are important steps and should not be overlooked. However, the principles proposed by these bodies are just that – principles. As is argued in this paper, there is a need for more entrenched regulations around the development of these systems, and these regulations need to be properly weighted with ethical consideration through the application of an appropriate ethical framework, like that which is proposed in this thesis.

[2] British company Cambridge Analytica and American company Facebook, to name two examples, are companies that have faced criticism and scrutiny over the sharing of personal data with external companies, specifically linked with the 2016 election campaign (See Granville, Kevin. "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens." *The New York Times*, 14 May 2018").

the EU, as will be discussed in Chapter 5 of this thesis, are currently developing regulations to mitigate these risks.

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, for better or for worse. Advancements in 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 due to existing technology (Lamb, 5). It is also estimated that the jobs most likely to be lost will be those belonging to lower-income and less-educated populations (Lamb, 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, 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.

The need for regulations in data ethics is a hotly debated issue in many capitals around the world. And there are several possible reasons that a professional code of ethics has not yet been created. As will be introduced by Paula Boddington, one reason is the threat that regulation can pose to creativity. The fact that development has primarily been with private, and therefore, less-regulated and less-scrutinized companies, is another. In fact, despite its rapid growth and spreading societal omnipotence and influence, only now are governments beginning to use these

technologies. This broadened application further strengthens the need for a professional code of ethics for the development of AI and automated decision-making systems.

Paula Boddington's book, *Towards a Code of Ethics in Artificial Intelligence*, provides an overview of what professional ethics are, where they came from, how they are applied and the dangers we may face of not developing professional ethics. Boddington demonstrates that codes of ethics, such as professional and leadership ethics, are important in that they "aim to mitigate the potentially deleterious effects, or the misuse, of such professional power" (39). She insists that codes of ethics can only function effectively with both adequate institutional and societal backing.

Boddington examines commonalities between the different reasons and approaches to the development of professional codes of ethics, listing them as being (among others): being regulators of relations between professionals, clients and others; that they ideally outline procedures for reporting problems and violations of codes; and that there is an assumption of professional value, where the assumed value also contributes to the relatively high social standing of members of recognized professions. She also demonstrates how codes of ethics (and laws and other regulations) have developed in response to catastrophes or scandals, recommending that codes need to be respective of cultural differences.

Further in her writing, Boddington reflects on the drawbacks to Codes of Ethics, asking "Can Codes of Ethics Make the Situation Worse? Yes" (53). She itemizes these drawbacks, stating issues such as: codes can create a separation of ethics from 'life'; the idea that ethics are someone else's responsibility to monitor and check instead of society as a whole, as in to "do the ethics"; the concept of 'work to rule', whereby the individual only feels obliged to work to the code of ethics and not beyond it; and, the danger that a code of ethics might actually be serving

the purpose of calming public anxiety, without actually making a difference to the substance of warranted public concerns. Further, she admits that, “there are indeed very hard questions about how to translate institutional ethical policies into practice.” According to Boddington, “A code of ethics might worsen a situation by tying the hands of professionals whilst those outside the profession can carry on a practice with impunity” (54-5).

This is not to say that a code of ethics couldn’t and shouldn’t be designed and implemented. Instead, it draws attention to the shortcomings of past practices in ethical code development — and the need to account for them in the code’s development and application.

An underpinning of ethics in the creation of these systems would benefit not only society and those affected by the programs but also the programmers and creators themselves through the establishment of a chain of responsibility. In addition to this, the concepts of professional behaviours and societal effects are joined by a third factor when discussing AI: the behaviour of the machines created (Boddington, 59).

> There may be different expectations for human and machine agency which are present but not fully articulated. This can have concrete and deleterious impacts upon any ethical conclusions which are drawn. AI is used to enhance or replace human agency. This means we must pay attention to questions about the boundaries of human agency and ‘normal’ human functioning. There needs to be careful consideration of different cases, given the varying nature of AI. The impacts of AI may not be just on its immediate use, but further afield within complex social systems, and careful attention should be paid to this (Boddington, 85).

Through the case study later in this thesis, these concepts of regulation and the protection of human agency in a world of technological advancements will be revisited. AI and ADMSs will have dramatic effects not only on the individual level but on society as well, which will be evidenced through the case study in Chapter 3 of this thesis. These are key components, I argue, as to why an ethical framework needs to be created and applied to automated decision-making systems and their off-spring. As clearly articulated by Boddington, the immediate use of these systems is one thing, the unwieldly and unregulated future of intricacies between human and machine is another. While a code of ethics is useful, the limitations around the intricacies and logistic limitations of codes can be addressed through a broader ethical framework, against which to apply new ethical considerations as they arise. Through the application of an ethical framework, I will demonstrate that we, as a society and as individuals, need to develop the regulations and policies through a broader ethical understanding that can protect now us all — and be further adapted for later.

2. BLACK BOX ALGORITHMS

a. Introduction and review of concepts

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 chapter, "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, 316)

In this scenario, we see compounding points of contention in the evolution and use of AI and ADMSs. We see the use of algorithms to make decisions, replacing (or some would argue enhancing) human jobs — but what does this mean for the implementation of these systems in our societies? We see the existence of bias as a very big issue with the use of these kinds of systems. Can AI and similar systems actually remove bias? We see the issue of responsibility of

the decision made, where the bank declares innocence even though they are the owners and operators of the system. What does this mean for the responsibility of outcomes? And what does that mean for the systems currently being used and the ones to come?

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 where 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 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 (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's 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, a professor who lectures and writes on 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" (107). Recent studies examining automated decision-making systems, as outlined in the next sections of this thesis, 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.

b. Current applications:

i) As will be demonstrated through the following three 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. As simply put by Pasquale, "Pattern recognition is the name of the game— connecting the dots of past behaviour to predict the future" (20). COMPAS

Several studies have been conducted examining and reviewing the results of automated decision-making systems and their effects. One example is the 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 *et al.*). 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 *et al.*). Specifically, ProPublica found that:

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. […] black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent);
- White defendants were often predicted to be less risky than they were. […] white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent);
- Even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants;
- Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63

percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists;

- The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants (Larson).

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" (Larson *et al.*). It is important to recognize 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" (Larson *et al.*).

One person affected by the biases of the COMPAS system is Eric Loomis. In 2013, he was arrested and charged with a crime that did not require prison time. However, when his case was reviewed by COMPAS, the judge denied probation, based on the outcome generated by the system's algorithm. Instead, the judge handed Mr. Loomis an 11-year sentence. Mr. Loomis, "challenged the use of the algorithm, as a violation of his due process rights to be sentenced individually and without consideration of impermissible factors like gender" (Israni). His challenge was rejected by the Wisconsin Supreme Court, which held that Loomis' due process rights were not violated by the algorithm's risk assessment, "even though the methodology used to produce the assessment was disclosed neither to the court nor to the defendant" (State v. Loomis). Subsequently, the US Supreme Court declined to hear his case, meaning, "a majority of justices effectively condoned the algorithm's use" (Israni).

The issue here is not only the use of the algorithm, but the abdication of responsibility by judges in rendering sentences and raises questions around which has a clearer or more detrimental bias: humans or the algorithms they build? While the human bias of judges are 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 the most well-meaning humans. But shifting the sentencing responsibility to a computer does not necessarily eliminate bias; it delegates and often compounds it" (Israni). 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).

As will be discussed later in this thesis, 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'" (41). This stigmatization is problematic for those with already increased vulnerabilities, as will be discussed in Chapter 4.

ii) 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, 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'Neil, 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, 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, 87). The problem is that these crimes, labelled as 'nuisance crimes' are more likely to occur in impoverished neighbourhoods. Further, without focusing on more major 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 further arrests in these communities. 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 is spoken for louder than we can speak, and it is spoken for on its own terms" (Cheney-Lippold, 488).

This feedback loop reaffirms the 'problem areas' in a city and reconfirms the negative connotations of a neighbourhood. In this way then, as we will see later in this thesis through the case study and discussion on vulnerable populations, 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[3]. 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

[3] Examples of this can be found in jurisdictions within and outside the US. Women and other vulnerable populations are even more likely to be subjected to incarceration for 'victimless' crimes and are likely to be victims themselves. "Many of the issues that contribute to imprisonment are interlinked. [In the UK], a high proportion of […] women have been victims of crime themselves – more than half of those given prison time in the UK in 2017, which may even be an underestimate. There was a similar trend in the US. This has led some researchers to call out that imprisoning these women is "victimising the victimised", especially considering that in many cases, the crimes against them were worse than what they were sent to prison for" (Hogenboom, Melissa. *Locked up and Vulnerable: When Prison Makes Things Worse*. http://www.bbc.com/future/story/20180411-locked-up-and-vulnerable-when-prison-makes-things-worse. Accessed 27 Apr. 2019).

is anything but. In our largely segregated [US] cities, geography is a highly effective proxy for race" (O'Neill, 86-7).

iii) 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" ("Dominated by Men").

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" ("Dominated by Men"). 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" ("Dominated by Men").

c. Ethical Considerations

The danger with black box algorithms is clear: decisions and outcomes can result from embedded biases. This danger is further exacerbated with 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" (32). While there are many positive effects that these examples can point to, 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.

Even if it could be argued or found that this discrimination would be evenly felt across the social spectrum, by random chance, for example, then perhaps it could be argued that these ends may justify the means. However, even if a person with fewer vulnerabilities were to be subjected to this form of erroneous discrimination, they would also, by virtue of their social

status, have more access and abilities to refute the charges or the bias against them. They would have also an increased ability to reclaim their autonomy and reduce their subjugation to further marginalization, perhaps even receiving compensation. The reality is that this is just simply a less attainable option for those who are truly the most affected by the racism and bigotry embedded in algorithms, that are based on the sample data provided to it by an already bias system, our own society. 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).

Are ADMSs in fact society's attempt to abdicate its responsibility to control for bias? "Bias can embed itself in other self-reinforcing cycles based on ostensibly "objective" data" (Pasquale, 41). There is great danger in promoting and assuming a system's benefits without regarding and accounting for its flaws. As seen in the COMPAS, PredPol and Amazon examples, there is an allowance for state-condoned rights infringements without fully understanding the background or context of a system. When society relinquishes the oversight of companies to account for and control for bias, it abdicates its ethical obligations and puts its future in those same hands without appropriate forms of accountability and recourse. We, as this society, run a great risk when we do not ascertain from an early stage the roles the companies creating these systems can play.

In *Society, Ethics and Technology*, by Morton Winston, philosophy educator and human rights scholar, and Ralph Edelbach, an associate professor in technological studies, they highlight the ripple effects surrounding this relinquishment:

> Often, one finds it impossible to attribute final responsibility to a single person or to a defined social entity. In the absence of a definition and a precise analysis of the responsibility chain, technologically advanced societies have shifted the issue onto the concept of risk assessment, thereby attributing value to the damage produced by entities that are seemingly devoid of responsibility… The question is: Who is responsible for any damage that may be caused by an autonomous robot? Is it the designer, the manufacturer, programmer, or final user? Often, it will be difficult to obtain easy answers to this question (Winston, 227).

3. CASE STUDY: THE BRIEF MENTAL HEALTH SCREENER APPLICATION (BMHS)

a. Introduction of BMHS: How data collection is being used to 'improve' cross-sectoral access and response to health for individuals

In this case study and by means of a vignette, I will review a program currently being implemented in Ontario, the Brief Mental Health Screener (BMHS) by HealthIM, as an example of an emerging system designed for use with vulnerable populations. This case study will be used to demonstrate the ethical considerations of vulnerability, autonomy and consent, further discussed later in Chapter 4 of this thesis, that need to be considered when collecting, using and disclosing the data of vulnerable populations within and between sectors. To do so, I will introduce and give the history of the program, the context around its development and adoption by the Ontario provincial government. I will then introduce the issues involved in the application through a vignette depicting an example of a scenario in which the application would be used.

As stated on the website of the application developer, HealthIM, "the BMHS was developed by Dr. Ron Hoffman in 2013 and is the first and only brief assessment tool purpose-built for use in law enforcement environments. Developed from a database of over 40,000 mental health assessments the BMHS uses the indicators most commonly associated with violent outcomes" (HealthIM (1)).

In Hoffman *et al's* article (2016), "The use of a brief mental health screener to enhance the ability of police officers to identify persons with serious mental disorders", the authors introduce the InterRAI Brief Mental Health Screener (BMHS). This screener is a paper-based version of the digital application discussed later in this case study, developed through an international not-for-profit consortium of researchers. The BMHS was originally created in

response to the criticism police offices had received, "over how they have responded to PSMD[4] [persons with a serious mental disorder] experiencing a mental health crisis particularly when there are tragic consequences. Two issues have been inevitably raised in the aftermath of such incidents: the [in]adequacy of police training on mental health issues and the recognition that there needs to be more effective integration and collaboration between the mental health and criminal justice systems" (Hoffman *et al*, 29). As explained in the article, prior to the study, when an officer was called in to a situation involving a PSMD, the priority had been ensuring the safety of the officers rather than the officer's ability to de-escalate a situation and meet the needs of the people with mental illnesses.

Hoffman *et al.*'s research paper reviews the study surrounding the origins and testing of the BMHS through a pilot project based in a community in Southern Ontario, "to describe the process that lead to the creation of the InterRAI BMHS, the findings of the pilot study and to explore the implications of its continued use" (Hoffman *et al*, 29). Say the authors, "police officers tend to focus primarily on violence potential and issues of public safety", due to the assertion that "current police training on the subject of mental illness is inadequate and lacks standardization across the country" (Hoffman *et al*, 32). The convergence of these factors has led to a number of deadly encounters between police and PSMD as well as to the recognition that there needs to be "more effective integration of systems of services" (Hoffman, 29).

The concept of the BMHS stems from the legal obligations faced by police under provincial mental health acts. These acts provide the authority for police officers to, "apprehend a person who they believe has a mental health disorder and poses a threat to themselves or others

[4] At the outset and of note, the authors used the acronym PSMD throughout the article. I will also do so as it relates to their article. However, the use of the acronym is in itself interesting in the context of this thesis, as a term they use to describe a human being in a time of crisis. This will important to reflect on in Chapter 4 of this thesis on vulnerability.

and to transport the person to a hospital for assessment or treatment, […] when they believe a mental health problem is present, there is a risk of harm to self or others, or there is evidence the person is incapable of caring for him or herself" (Hoffman *et al,* 28).

Hoffman *et al.* introduce the InterRAI BMHS as a "new evidence-based brief mental health screening form developed for use by police officer", with two main goals: 1) "to enhance the ability of police officers to recognize indicators of serious mental disorder" and 2) "to help facilitate a more collaborative relationship between frontline staff in the criminal justice and health systems" (Hoffman *et al,* 29). Historically, the relationship between these two frontline systems has been "adversarial", according to the authors, due to the difference in their respective areas of focus, with police focusing on public and personal safety and the emergency departments focusing on the patient, the severity of their symptoms and needs while in hospital.

The authors also point to the fact that screening and rating instruments related to mental health have been used for over half a century, citing examples such as the Brief Psychiatric Rating Scale, Beck's Suicidal Intent Scale and Beck's Hopelessness Scale. The InterRAI BMHS is a necessary next step, say the authors, "given that none of the mental health screeners currently in use are suitable for frontline police officers and ED [emergency department] staff" (Hoffman *et al,* 30).

They found that the results of the BMHS pilot study are "consistent with the literature in that police officers tend to focus primarily on violence potential and issues of public safety" (Hoffman *et al,* 32). Say the authors, "the use of the InterRAI BMHS enabled the collection of more detailed information on the characteristics of persons who police officers had interaction with," and "enabled police officers to capture, in much more detail, the characteristics of PSMD who were hospitalized but were not based on diagnoses" (Hoffman *et al,* 32).

Most importantly for this thesis, the authors acknowledged the limitation of the study. Limitations included factors such as the relatively small sample size and that the sample was, "biased toward [a] white, predominantly English-speaking population with almost no representation from First Nations peoples" that "does not reflect the demographic reality in the province, particularly in larger urban centres" (Hoffman *et al*, 33). Further, there was an inability to gather randomized or matched control sample of police interactions with PSMD from other parts of the provinces and it was not possible to measure inter-rater reliability or to estimate the validity. Lastly, the authors cite that the "dependent variables were limited to police escort to hospital and admission, which does not reflect the current broad range of dispositions available to both police officers and clinicians" (Hoffman *et al*, 33). Hoffman *et al.* recommend that future research should work to address these limitations of inter-rater reliability and the integration of cross-validation into the research design as well as determining if the officers benefited from the enhanced training they received in using the BMHS form itself.

Overall, the authors conclude that the use of the InterRAI BMHS did "enhance the ability of police officers to identify indicators of serious mental disorder" and that the pilot study "provides useful information for both police officers and clinicians regarding variables that are significantly associated with serious mental disorders" (Hoffman *et al*, 33). The study also references the creation of electronic applications to, "further enhance communication all of which is ultimately aimed at better addressing the needs of persons with serious mental disorder" (Hoffman *et al*, 33). As will be demonstrated, the "electronic applications" Hoffman *et al.* reference here are central to the discussions and analysis contained in this thesis.

In 2015, the Ontario government's Ministry of Health and Long-Term Care, under the advice of the Ontario Health Innovation Council (OHIC), created and appointed the new position

and office of the Chief Health and Innovation Strategist. This position was created to, “champion Ontario's health technology innovation sector” (Ontario, OCHIS). Along with the new position, the OHIC also recommended: “establishing a new $20-million Health Technology Innovation Evaluation Fund to support made-in-Ontario technologies; using newly created Innovation Broker positions to connect innovators and researchers with opportunities in the health care system; streamlining the adoption of health care innovations across the health system; and, shifting to procurement practices that focus on outcomes, such as fewer hospital readmissions and the long-term value of medical devices” (Ontario, OCHIS).

It was through these recommendations that new funding became available for advances in health technologies. Said Dr. Eric Hoskins, then the Minister of Health and Long-Term Care, “[the investment of $20 million in health research, called the Health Technologies Fund (HTF)] recognizes the depth of innovation that exists here in Ontario, and its role both in transforming our health care system and in improving the patient experience. With the support of Ontario’s new Office of the Chief Health Innovation Strategist, those ideas and inventions will make their way into our health care system—be it a hospital, a doctor’s office, or a long-term care home—to change the way we deliver care for the benefit of patients. It’s a win for Ontario’s patients, a win for the health care system, and a win for Ontario’s economy” (Ontario, OCHIS).

The promotion of new technologies in health care and access is a focus for the Office of the Chief Health and Innovation Strategist (OCHIS), with targeted priorities to procure innovations, “by shifting the health care system to strategic, value-based procurement and removing barriers for small and medium-sized enterprises to participate” (Ontario, OCHIS). The ‘value-based approach’ is focused on new ways in procuring and providing services to Ontarians, promoted as a modern approach to meeting the demands of a single-payer system. It is promoted

as a cross-sectoral approach to health care delivery, “sharing the risks and benefits of new technologies and processes with industry and health service providers” (Ontario, OCHIS). The focus is on the improvement on service delivery for health care providers and technology providers all with the proclaimed goal of providing better access for Ontarians. In 2016-17, $5.4 million of the $20 million HTF was provided to 15 innovative health technology firms, examples of which include “SMArTVIEW, a program that provides monitoring and self-management for patients following cardiac and vascular surgery” and “Intelligent Scheduling to Reduce MRI and CT Wait Times”. Most of the funding was applied to new platforms for better data collection, to improve and report on access for Ontarians to health care services. For this case study, I will be focusing on one of the innovative technologies that received funding from the HTF, “Improving Mental Health by Connecting Police and Community Services” (Ontario, OCHIS).

b. Introduction of the BMHS automated decision-making application

Of the 15 projects funded, the item entitled “Improving Mental Health by Connecting Police and Community Services” received a grant of $498,000 to create software to “facilitate assessment, risk appraisal and case management of individuals with serious mental disorders” (Ontario, OCHIS). The software provides a link between first responders and other health and community services in a time of crisis for an individual who is affected by a mental health issue. HealthIM was one of the recipients of the funds, with which they created the Brief Mental Health Screener (BMHS) Application. HealthIM is a software company founded by University of Waterloo graduates that works to assist and support police services when interacting with individuals in crisis due to mental health issues. As described on its website, “HealthIM is dedicated to building stronger communities by providing a high-quality and meaningful software platform to police officers. Every police agency we support is devoted to supporting their

community and each member of our team is committed to helping these efforts" (HealthIM (1)).

The BMHS is an application housed on an officer's smart phone or laptop that helps the officer to approach and assess an individual in a mental health crisis with clinical assessments and language. There are four features of the application:

> 1) a digitized screener, which assists in an evidence-based mental health assessment, provides a digital officer interface while optimizing the users experience and employs algorithms to predict the risk of harm that an individual may pose;
>
> 2) a wireless connection to a hospital, in addition to providing the hospital with a clinical summary of the individual and a history of previous encounters between law enforcement and the individual;
>
> 3) paperless record keeping, which digitally stores records, transmits report data to police and health care partners, as well as the ability to integrate with other, common records-management providers in Canada; and,
>
> 4) an analytics dashboard, providing real-time access to data analytics and metrics, automatic report generation and internal performance management metrics for police services (HealthIM (1)).

A short video is posted on the HealthIM website depicting an encounter with the Mobile Crisis Rapid Response Team (MCRRT), a pilot project in Brantford, ON, which partners a police officer with a mental health worker to be dispatched together to respond to calls involving a

person in crisis[5]. The program cites its success, stating that, "this is the first program of its kind in Canada, [it] has reduced our apprehension rate from 66% to 25% and people are getting the quality health care they need in a timelier manner" (Mobile Rapid Response Team). The video depicts an interaction that is an example of this community support program for individuals with mental health issues. For this thesis, I will use it as an example of an interaction where the BMHS application is likely be used (HealthIM (2)):

> The police officer involved in the demonstration begins by stating there is an identified need for this program in their community as a result of a "high-volume of calls" across various front-line services. As part of the Mobile Crisis Rapid Response Team (MCRRT), the initiative was brought forward to "reduce the number of calls and burden on emergency services and the program seems to be effective in doing that by making connections with people in the community". The video also introduces a Mental Health Specialist alongside the police officer, who works in the community and is now able to conduct a mental health exam "on the spot" to decide what further treatment they will need – either their own doctor or a hospital. The demonstration then introduces a "member of the community that is in crisis".
>
> In the demonstration, a man is sitting on the front stoop of a house. It is a cold day and he is rocking back and forth, appearing distraught about something. The officer and the Mental Health Specialist approach the man and introduce themselves, asking him how he is doing. When asked by the officer, he says he has lost his job and, "now is not a

[5] More information on this program can be found at: https://www.stjoes.ca/health-services/mental-health-addiction-services/mental-health-services/coast/mcrrt (MCRRT)

good time for any of this to be happening". The officer asks him his name and the man appears to hesitate, starting to ask, "Is there…?" and then responding, "Shawn."

Shawn gives some details about how he lost his job and that now he has nothing. He asks, "Am I in trouble?" The officer introduces himself and informs Shawn that they go from "call to call, talking to people that are in crisis, just like you appear to be in right now. We're here to help." They then go on to say, "Now, crisis doesn't mean that you have a mental health problem, crisis means that you are having a rough time with life right now, and you need someone to talk through things, and that's what we do." Listing the access to services they can provide, Shawn replies that he doesn't really think they can help him. He then states he would be better off alone. The Officer replies, "Well Shawn, right now I'm with you and I think you need us as much as anybody else in this city does." Shawn appears embarrassed, saying he, "shouldn't be acting like this." "It's perfectly fine," replies the officer. "What would happen if we left you alone right now?" Upset, Shawn says, "Everybody else leaves me alone […] so you wouldn't be the first, so it doesn't matter, man. I'm sorry — officer." The officer responds that he is going to stay with him "as long as it takes" to help him through the problems he is having. "But you've got to give us a little something alright? You've got to talk to us and we can help you work through it, okay? Are you willing to do that for us?" "Okay", replies Shawn, "but it's not like I need an ambulance." "No," they reply, "we're just here to talk to you […] you're not doing anything illegal and we're not here to arrest you, we're trying to avoid having anybody arrested or go to the hospital or anything like that. That's not what we're here for. We're here to talk with you, we're going to stay here as long as it takes. We're here on your front porch and might be a little more comfortable inside, or if you're okay

out here…" Replies Shawn, "Well, if you want, we can go inside because my neighbours are always looking anyways. They're probably the ones who called you." Says the Mental Health Specialist, "Someone called us, we're not really too sure who called, but they called because they're concerned." Dismissively, Shawn replies "Whatever, I, yeah, I mean, if you guys want to come in, it's fine, I mean, I don't know how much time you have, like I said I just don't want to take up your day." "We've got all day, you're our number one concern at this point, Shawn," replies the Officer. "Alright," agrees Shawn, "Come on in then."

In a 'debrief' after the video, the Officer and Mental Health Specialist determined that the individual was in a more situational form of crisis, needing only referrals to community support and information, "helping him focus on his problems and getting him through some of the issues at hand to prevent the crisis from coming up again." Says the Mental Health Specialist, "You're trying to break the stigma of mental health. Everybody has situations in their lives they're not able to handle by themselves and they need further assistance in helping themselves in the future." The Officer states that follow-up with the individual in crisis occurs within 48 hours in order to check on them and "ensure that they are on the right path to avoid crisis in the future," which often surprises individuals. "In doing this," affirms the Officer, "we ensure that nobody is left behind, nobody is left out, nobody is forgotten." The video then closes with statistics, stating that, "Since its inception in September 2015, the Mobile Crisis Rapid Response Team has successfully diverted 74% of individuals in crisis away from the healthcare system."

While the video does not specifically depict the BMHS application being used by the officer, this is a situation in which the application would be a tool for the officer to use to determine the risk of the individual to themselves or those around them. The central issue here is one of consent. There is no clear understanding that consent would need to be received from the individual prior to the BMHS application being used to collect and analyse their data. This is a common situation faced by individuals when dealing with consent and persons with mental health issues or other vulnerabilities, as will be discussed in Chapter 4. Further, there is a power dynamic at play here, where the officer is not just a friend checking up on Shawn, but rather a person with great authority and exercisable power. Would Shawn feel that he has the ability to say "no" to the Officer, either in answering his questions, allowing him to enter his house or collect his information for use in the BMHS application? Does Shawn have the ability to understand the effects this situation and the answers he gives to the Officer might have on him in the present, near or longer-term future? Would Shawn have been treated differently if, for example, the crisis was not that he had lost his job but the he was unemployed to begin with, if he lived in a more affluent neighbourhood or if he had a family or other support network of people who have been able to interject, ask questions on his behalf, or merely be present to witness this interaction? And finally, just because Shawn is in crisis now, does that imply that the data collected on him in that moment can, and should, be collected, stored and shared without his knowledge in the future, even when he may no longer be in that state of crisis? Questions such as these are raised through interactions such as the one between Shawn and the officers, but cannot be properly answered by reducing Shawn to his data and any subsequent statistical analysis. This vignette also raises the following three questions, to be reflected on throughout this thesis: 1) Is consent sought from the "PSMD" to have personal information collected and later used by the

BMHS system? 2) How is the privacy of the "PSMD's" personal information protected and who has access to this personal information?, and, ; 3) Where is this personal information stored and how is it accessed beyond the BMHS application, and what are the future uses of it beyond the intent through which it was originally collected.

The collection, storage and disclosure of data has many ethical implications that must be considered when dealing with any members of society, particularly those vulnerable members of society who have fewer avenues of recourse and are among the most frequent beneficiaries of social and government-provided services. The BMHS is an example of a source of data-collection with direct access to those more vulnerable members of society — and, even more pointedly, while in their most vulnerable states. Through the analysis of my case study in Chapter 6, I will review the BMHS and examine how the RtE and the RtbF could be applied to strengthen autonomy for the individuals who come into contact with this system and its accompanying algorithms. Governments are working to embrace new technologies, both to meet the needs of their citizens and also to build on what is being developed in the private industry. There is much to be gained by citizens with a cross-sectoral response to their needs. The intersectionality of the social determinants of health is an example of this. The connection between mental health and social and physical health outcomes is a highly evidenced example of this, wherein the mental state of an individual has a direct effect on their social and physical well-being. However, the ethical implications in these connections through black box algorithms must be examined, clarified and addressed.

c. What societal impacts could this application have on vulnerable populations?

Compounding the concerns of the BMHS is its privacy policy. As a pilot project that was later implemented by local police forces, the emphasis of the privacy was on the subjects of the

study, the police officers themselves, not the "PSMDs" or citizens that the police are interacting with, while collecting and sharing their data and personal information. As outlined in the HealthIM privacy policy for the BMHS application:

> In order to use the services provided by HealthIM, you may be required to provide specific information about yourself, such as your name […] and about individuals whose information you wish to input into the services. We will always inform you when we need information that personally identifies you and we will keep this information in strict confidence – however, by using HealthIM's services, you acknowledge that you will be providing HealthIM with information about individuals that they may consider to be their personal information. It is your responsibility to make sure that you have the right to disclose their personal information to HealthIM, and to let them know that their personal information will be handled in accordance with this privacy policy (HealthIM (3)).

In this document, it appears that the privacy policy is a contract only between the officers and HealthIM and pertaining to their personal data as priority over the data of the "PSMDs" collected. Other problematic portions of the privacy policy include statements such as "HealthIM uses reasonable precautions to protect personal information and store it securely. Access to your personal information is restricted to those personnel who need it […] HealthIM only uses and discloses your personal information if it helps us facilitate your needs, improve our products and services, meet legal and regulatory requirements, or – to the limited extent possible – manage our internal business operations" (HealthIM (3)). Further, "HealthIM may combine your information with other information into an aggregate form, so your information no longer personally

identifies you. We may then disclose the aggregate information to third parties, so they can obtain an overall picture of HealthIM products, services, customer sectors and/or usage patterns" (HealthIM (3)). Additionally, HealthIM declares that it will only share an individual's personal information with their consent. However, it clarifies that:

> in a few situations, HealthIM may be required to disclose your information with your prior consent", citing necessity by law or if, they "believe in good faith that such disclosure is necessary to (eg. Protect and defend our rights and property, or the rights and property of a third party; protect the personal safety of any person; [or to] allow for a change of ownership of HealthIM and associated transfer of all personal information to the new owner of HealthIM – this does not affect the protection of your information under this Privacy Policy Health. HealthIM will always try to provide you with prior notice of such disclosure; however, such notice may not always be possible or reasonable given the circumstances (HealthIM (3)).

Interestingly, the privacy policy does address the destroying of information, which will be important for this thesis when discussing the RtbF. The policy states: "You can request that your personal information be destroyed at any time a situation may arise where you desire to have all of your personal information that is contained in HealthIM's records deleted or destroyed. […] However, there may be situations where we are obligated to retain one archival copy of your information to allow us to comply with laws or respond to legal processes. We wish to inform you of all such situations and will only use your retained personal information to the limited extent necessary to comply with such laws or respond to legal processes" (HealthIM (3)).

This will be referred to later in this thesis as problematic for its vagueness, specifically in the fact and as evidenced earlier, that it refers only to the information of the police officers collected through the application and not the data collected on the individuals they interact with through the BMHS application.

These items will be more closely reviewed, in Chapter 5 of this thesis, in the context of a feminist bioethical framework for vulnerable persons and then two government's legislations on privacy. The main points to keep in mind regarding the BMHS as a case study throughout the remainder of this thesis are those of the ability for an individual to give consent, or their inability to give consent, either because it is removed from them due to their mental state (discussed in Chapter 4) or because they are unable to truly agree to consent as they feel pressured to do so as a result of the power dynamic in interacting with police or other persons of authority. Further considerations include the issues of privacy around the access to the data collected, either at the time of collection or any time in the future, as discussed in Chapter 5. Lastly, it is important to consider who has the ability to share this data, even if beyond the original purpose of its collection, and how or where it will be stored, also examined in Chapter 5.

4. FEMINIST BIOETHICS, VULNERABLE POPULATIONS AND CONSENT – A NEED FOR A CONCEPT OF VULNERABILITY THAT INCLUDES AUTONOMY

a. Vulnerability and Autonomy

The previous chapters have looked at overall concepts and some of the ethical considerations around AI, ADMSs and black box algorithms, introducing a case study that will be revisited throughout this thesis. The generalized bias that is contained within ADMSs has been demonstrated and this raises concerns over how the privacy, autonomy of persons can be addressed. However, as has been discussed, vulnerable populations are at even greater risk in general and this is even further exacerbated with ADMSs. Their already diminished access to autonomy, reduced by their reliance on social services and their marginalized status, is further compounded when bias and paternalism is induced even more covertly through these systems. Furthermore, consent is an even more complex concept for these populations as they are exacerbated by the vulnerabilities of individuals.

In this chapter, I will review a feminist bioethical theory on vulnerable populations and consent. I will be using this framework for the case study in this paper as it best addresses the concerns of vulnerable populations and is also well-suited for potential adaptation towards an ethical framework for the collection and use of an individual's data. The framework works to ensure an individual's autonomy by examining the world in which they operate and the confounding factors against their ability to attain it, especially how vulnerability can be created and exacerbated by policies.

In Rogers *et al.*'s article, "Why bioethics needs a concept of vulnerability", the authors work to strengthen and reassert the concept of vulnerability in bioethics through a feminist lens

as important and yet often overlooked. They take a historical look at approaches to vulnerability in fields from research ethics and public health ethics and determine the common threads in these fields in their concepts of vulnerability and how these are addressed. The authors find and support four common areas of vulnerability that they assert need to be considered when dealing with vulnerable populations: harm, exploitation, needs and autonomy. Using the definition of vulnerability, as "being at increased risk of harm, and/or having decreased capacity to protect oneself from harm" (12), the authors hold that the concept is undertheorized to "assess or justify the interventions and practices invoked in the name of protecting the vulnerable."

Rogers *et al.* assert that vulnerability is an ontological part of the human condition (as in, as a result of being human, we are vulnerable), wherein "human life is conditioned by vulnerability" (12), and as a universal construct, wherein there are components that determine that one person can be at a greater disadvantage than another. As such, they state the need for, "identifying different sources of vulnerability and the different ways in which vulnerability is realized" (12), to ensure that society responds with the appropriate moral response. Without doing so, society runs the risk of doing too little, potentially "failing to recognize a source of vulnerability", or too much, therefore leading to an increased risk of "paternalistic protections" (12). Rogers *et al.* also cite the requirement that a concept of vulnerability not just protect from harm but that it also "attends to the ways in which the development of capacities for resilience and the social conditions for promoting agency and autonomy" are incorporated in applications of the concept (12).

Defending their search for a rooted and true concept of vulnerability, the authors state that, "developing a robust and nuanced account of vulnerability [in bioethics] is necessary to allow us to identify sources of vulnerability and to determine just who is vulnerable, be this at

the individual, group, or population level; to ground duties such as protection to those who are vulnerable; and to recognize the circumstances in which interventions to ameliorate are warranted" (13). Through the identification of vulnerabilities and an individual's exposure to them, the programs and policies designed to assist the individuals would work to only assist in those areas of their lives affected by these vulnerabilities. By reducing unnecessary external interventions into areas unaffected by vulnerabilities, their autonomy would be further supported and protected.

The authors cite the Belmont Report, which built on an early concept of bioethics stemming from the Nuremburg Code and the Declaration of Helsinki, which is applicable in that it focussed on human subjects. The Report cited three characteristics of vulnerability: 1) a lack of capacity to consent to research; 2) an increased susceptibility to consent to research; and, 3) an increased risk to harm. Through this 'labelling approach' to vulnerabilities and on these grounds, the Report stated that the vulnerable should not be included in research. Rogers *et al.* cite critiques of the social injustices stemming from this exclusion, outlining that without the inclusion of those labeled vulnerable, research findings would be devoid of the needs of, or applications to, these populations. Critics of the labeling approach stated that the approach was too narrow as it "reduces vulnerability to questions of competence" and therefore risks their exclusion and being "unduly paternalistic", and too broad (being "over-inclusive") to the point that "nearly everyone had been identified and labeled as "vulnerable". This results in "obscuring rather than enabling" their identification as needs (16).

One response to the 'labeling approach' is to apply 'analytical approaches', which have "attempted to identify sources of vulnerabilities in order to assess their impact and develop remedies" (16). However, Rogers *et al.* found these approaches, which are rooted in the research

ethics, to be inapplicable outside of a clinical-like setting. They found both the 'labeling approach' and 'analytical approaches' to be missing the concepts of autonomy, the demands of justice and their relations with vulnerability.

Turning to public health to see how it approaches and defines vulnerability, the authors find this approach to support a link to the demands of justice, where social determinants of health act as links between many various facets of an individual's life that can lead, or turn away from, a person's ability for social connection and relative success. In public health, vulnerable populations are defined as,

> those who, because of 'financial circumstances or place of residence; health, age or functional or developmental status; ability to communicate effectively; presence of chronic or terminal illness or disability; or personal characteristics,' are unable to safeguard their own needs and interests adequately (17).

Two sources of vulnerability are identified, which are overlapping in their effects: 1) as a "shorthand" or a "marker for disadvantage", for those who are deprived in one or more ways related to the social determinants of health and therefore at "higher risks of poor health"; and, 2) referring to those who already have a form of disadvantage and are therefore at greater risk for other forms of disadvantage and "further suffering" (17-18). While distinct in with its own understanding of vulnerability, the public health concept of vulnerability points to how these two sources, "often coexist and compound each other through complex interrelationships in vicious cycles" (18). Overall, public health approaches are primarily rooted in social justice, linking duty and commitment to those vulnerable populations. However, Rogers *et al.* point still to the lack of autonomy in addressing vulnerable populations in these approaches.

Applying a philosophical approach to vulnerability, the authors find that vulnerability is the result of humans existing -- simply being embodied. As a result of this embodiment, all humans have some level of dependency on others, which in turn reveals the capacity for all humans to be vulnerable to some degree. Following this, the authors point to the generation of obligations because of this universal vulnerability. Citing Goodin and his book, *Protecting the Vulnerable*, Rogers *et al.* draw on his concept of vulnerability as relational, pointing to Goodin's "principle of protecting the vulnerable" (PPV). This principle holds that the goal should be at least to not increase vulnerability, stating that, "we have an obligation to act so as to prevent harms to, or protect the interests of, those who are especially vulnerable to our actions or choices" (20). This, according to the principle, is as a result of moral obligation and consequentialist models due to Goodin's link with "correlative responsibilities" – "conceptual connections between vulnerability, harm and exploitation" (20-1). Rogers *et al.* give the responsibility of a parent as an example of this: where, "parents have special responsibilities to protect and promote their children's welfare because a child's present and future welfare are especially vulnerable to her parents' actions and choices" (20). PPV puts a responsibility, in the form of obligation, on the more powerful party to guard against their own abuse of power and to protect those vulnerable to their power.

The authors find the following two issues with Goodin's PPV: 1) Goodin claims that "the causal history of relationship is irrelevant in determining the responsibilities arising from vulnerability"; and, 2) PPV claims to "explain responsibilities arising from vulnerability in relationships involving inequality, but can also ground broad social welfare obligations" " (21). The authors find flaws with the PPV: 1) with the former, the relationship is not irrelevant but is in fact, "crucial in determining" responsibility for and response to the vulnerable; and, 2) for the

latter, if everyone is potentially vulnerable, this then makes it difficult to determine moral priorities in these relationships and damages the value of the concept of the vulnerability. They highlight the fact that if all are vulnerable to some degree, it ignores the need to give moral priority to some needs and correlative vulnerabilities – outside of "vital needs", say the authors, "those that are inescapable and without which the being in question will be seriously harmed or fail to flourish", of which even they are situationally dependent, and are given moral priority by Goodin and Needs theorists, such as Reader and Wiggins (22).

Needs theorists, according to Rogers *et al.*, work to bridge the gap between, "a universal claim about the normativity of needs […] with a context-sensitive analysis investigating specific needs claims that are inevitably social, culturally and politically contestable" (22). Following this view, vulnerability is the exposure and risk of serious harms to one's vital needs – "harms that impair one's ability to lead a flourishing life" (22). The fact that all humans have needs and, therefore, have the potential to be vulnerable to someone in order to meet these needs is measured against the inability of some humans to meet their vital needs. As a result, a reflection on Needs theories leads Rogers *et al.* to support the concept that vulnerability is "as a matter of degree" (22). However, the authors draw attention to the lack of autonomy and agency as a component of vulnerability in these theories and their conclusions, one that they deem to be conspicuously lacking and necessary. The risk in this is the tendency for "unwarranted paternalism" (23). Citing this, the authors seek a more pluralistic approach.

The authors introduce John O'Neill, who points to the frequency of humiliation as inconsistent with autonomy in concepts surrounding vulnerability. On this point, Rogers *et al.* seek to include autonomy into an analysis of vulnerability that, "explains why we have obligations not only to protect vulnerable persons from harm, but to do so in a way that

promotes, whenever possible, their capacities for autonomy" (23). In this respect, vulnerability as ontological – or the nature of being -- creates a "sense of solidarity" and "encourages responses to 'more than ordinary vulnerability'" (23). Drawing on this, the authors introduce the relational approach to autonomy, which highlights autonomy as a socially constituted capacity, wherein it is a product of many interrelated social webs. This approach highlights that the development of autonomy, "can be impaired and its exercise thwarted by exploitative or oppressive interpersonal relationships, and by repressive or unjust social and political institutions" (23). In essence, there is an "inescapable dependency on, and hence vulnerability to, others" (24). As a result, say Rogers *et al.*, it is the relational approach that supports the concept of an obligation towards the vulnerable needs to match the focus of protection with the promotion of autonomy for those who are, "more than ordinarily vulnerable" (24).

Following the juxtaposition of these theories, the authors provide three taxonomies of overlapping, yet different, kinds of vulnerability: 1) inherent vulnerability, which arises from our embodiment as a result of our "neediness" and dependency on others; 2) situational vulnerability, which is context specific and, "may be short term, intermittent or enduring"; and, 3) dispositional or occurrent vulnerability, where dispositional is by virtue of existence (e.g., humans need food to survive and therefore are all vulnerable to hunger), where occurrent is as a result of our location, access, etc. (e.g. some humans lack access to adequate food and nutrition and are therefore occurently vulnerable to hunger) (24-5). Both inherent and situational vulnerability can be either dispositional or occurrent vulnerabilities. Say Rogers *et al.*, responses to inherent and situational vulnerabilities usually aim to support those who are occurently vulnerable and reduce risks of individuals and populations to avoid occurrent vulnerabilities. Rogers *et al.*, applying the aforementioned theories and taxonomies, argue that all interventions for vulnerable populations

must, “enable or restore the agency” of these individuals and that, “this is most likely to be achieved by interventions that engage their agency and participation, wherever possible and to the greatest extent possible” (25). The authors also introduce pathogenic vulnerabilities as those that are generated by morally dysfunctional interpersonal and social relationships and/or policy that addresses vulnerability while still increasing vulnerability. Say the authors, all of these vulnerabilities can be linked in that, “they can engender a troubling sense of powerlessness, loss of control, or loss of agency. Our emphasis upon obligations to foster or restore agency wherever possible motivated by the recognition of [pathogenic vulnerabilities]” (25).

In reflecting on examples of international social policies for vulnerable populations, Rogers *et al.* demonstrate how, “within social policy contexts, the identification of a person of group as ‘vulnerable’ can lead to unjustified paternalism (‘labeling’); and the ways in which policies intended to support the vulnerable can lead to what we have described as ‘pathogenic vulnerability’” (26). Assert Rogers *et al.*, “A well-articulated vulnerabilities approach could better assist in identifying the complex relationships among culture, historical injustice, social disadvantage […] that are relevant here. Such an approach would carefully articulate the different kinds of vulnerability involved: the inherent vulnerability of [individuals], the situational features leading to occurrent vulnerability, the presence of pathogenic vulnerability, and the needs and harms involved” (28).

b. Autonomy, Consent and Voluntariness

Further examining vulnerability and autonomy through a feminist bioethical lens, in "Competence, Voluntariness, and Oppressive Socialization: A Feminist Critique of the Threshold Elements of Informed Consent", Dominic Sisti and Joseph Stramondo add to the importance of autonomy the concept of oppressive social norms and its effect on the ability to grant informed consent by individuals. As will be seen, their work supports the importance of autonomy and its necessary inclusion in an ethical framework for persons labeled as vulnerable, which I will later apply in this thesis to ethics around data collection and use.

Sisti and Stramondo work to demonstrate how the standard bioethical model for informed consent is, "inadequate because it relies on presumptions of procedural autonomy and rational choice that overlook the problem of how agents are often socialized so that they adopt and internalize oppressive norms as part of their motivational structure" (67). Referring to Tom Beauchamp and James Childress' 2013 paper on the thresholds for informed consent — namely, competence and voluntariness — Sisti and Stramondo take exception to these thresholds. In their paper, they bring to light and clearly demonstrate how these thresholds in themselves are problematic and need to be reworked (68) in order to provide access to true autonomy and autonomous decisions by oppressed persons.

As a feminist critique, the authors reveal that for some feminist ethical theorists, this is satisfactorily addressed through procedural or process-based models of autonomy, wherein a "checklist" of sorts is created to ensure that the individual's autonomy is supported through the consent process. This process stipulates that certain conditions must be met for a choice to be considered rational. However, for the authors, this is problematic in that it dissolves the individual and their autonomy into parts and pieces and does not work to the keep the person as a

whole during the process. The resulting breakdown allows for normative content, wherein social oppressive can permeate the process. Sisti and Stramondo point to the fact that one can be, "procedurally autonomous while being substantively nonautonomous" (71), due to, as we will see, an individual's lack of awareness of the oppressive socialization they have been exposed to. This implies that while the checklist of autonomy can be 'checked' or confirmed, an individual's autonomy can still be at risk because of unaccounted for or underlying factors.

The standard bioethical model for informed consent, according to the authors, has three components: 1) the thresholds of competence and voluntariness; 2) information elements; and, 3) elements of the consent itself (69). Information elements refer to the threshold that the subject understand both the information provided and the planned use of their consent. Elements of the consent itself refers to the individual's decision in favour of the plan and the resulting authorization (or lack) of the plan.

It is with the first component, the thresholds of competence and voluntariness, that the authors take the greatest issue. They use the definition of competence as, "the ability to perform a task" (72) by Tom Beauchamp and James Childress, both renowned philosophers and creators of an ethical framework on biomedical ethics. To this, Sisti and Stramondo point to the "circular construction of competence", wherein "competence is framed as the ability to make an autonomous choice, while autonomy is determined by an assessment of a person's competency" (72). This circular reasoning is evidence of the socialized oppression, which "fails to encompass subtler forms of competence or, more important, incompetence" (73). The problematic circle continues and the implications are (at least) two-fold: if the competence to give consent exists (according to societal norms), then this does not protect an individual from being poorly

informed prior to the consent given; and if competence doesn't exist (according to societal norms), then that individual's right to give consent or have it required from them is taken away.

Sisti and Stramondo point to adaptive preference formations, introduced by Jon Elster, wherein an individual's preference is shaped by what they see as being available to them. This concept is rejected by some feminist theorists, wherein adaptive preferences are autonomy deficits. Those opposed feel that it, "'discourages us from treating people with adaptive preference as the types of people who can make authoritative decisions about their own lives... Characterizing people with adaptive preferences as incapable of choice leads us toward seeing people with adaptive preference as appropriate objects of coercion'" (74).The authors point to the disconnect between a person's inability to make choices based on their experience and instead to, "'choose and care about things that are not consistent with their flourishing.'" (75).

Following this, the authors argue for a normative theory of competence which offers normative content that includes an individual's motivations, "as not merely the product of oppressive socialization" (75). This would work to determine whether a person's competence has been "diminished by oppressive socialization and to examine critically the effect of such socialization, [to ask the question] "Do these motivations unjustly subordinate this individual's well-being to the unfair benefit of another with a more privileged social status?" (75).

Gaslighting is used an example by Sisti and Stramondo, following research wherein a person's ability to understand the reality of a situation is altered through emotional and/or psychological abuse at the hands of a perpetrator. "Women and others who have been socialized to adopt oppressive norms often question their own moral authority to make competent decisions in ways that those who have not been socialized in this way do not […] A person could have been gaslighted by her internalization of oppressive norms that imply she is not a competent

moral agent who can make an adequate choice about a course of treatment" (76). As seen earlier in this chapter, this echoes the same sentiments of autonomy and vulnerability as Rogers *et al.* As we will also see later in this thesis through discussions and analysis around the case study, this reality is directly avoided and ignored in dealing with people who are vulnerable and victims of oppressive socialization. Sisti and Stramondo write: "shame in oneself, rooted in social expectations, is enough to erode an agent's sense of self-worth for normative competence" (76).

On the point of voluntariness, the authors acknowledge that while it does "work to ensure that blatant forms of oppression or coercion are not at work, we believe the threshold element of voluntariness fails in cases of more insidious forms of oppression" (77). The authors point to the problem of power dynamics as they pertain to an individual's ability to be 'voluntary' in a situation where they are being/are oppressed: "voluntariness is preoccupied with [the] most obvious power relations. This is why is does not seem to be equipped to deal with our problem of oppressive socialization" (77). Turning to Benson and "free agency", the issue surrounding socially-unacceptable behaviours — such as "drug dependency, psychological illness, or profoundly oppressive social conditions" — alter the conversation about "what it is to act freely" (79). True voluntariness is obstructed by obvious cases where it is problematic, leading to the avoidance of theoretical assertion or "sensitivity to detect subtle, socially formed oppression" (79).

As ways forward, Sisti and Stramondo propose the re-examination of these threshold elements, which "would strike a balance between having enough normative content to recognize when a decision is being unduly influenced by oppressive socialization and not having so much normative content that decisions are deemed nonconsensual and invalid if they do not align with certain normative standards" (81). The authors itemize the potential in revised models of

autonomy by liberal-cultural feminists, such as Atkins and Meyers (81) who, with a relational model of autonomy for nursing, find solutions through dialogue with patients, wherein the patient is helped to find the, "words and descriptions needed to articulate beliefs, emotions, desires, and values [and that] helps the patient examine, evaluate, and prioritize these phenomena [that would help] the patient bring [these] together into a point of view" (qtd. in Sisti and Stramondo, 81). This, say the authors, engages the patient and, at the very least, has the potential to raise their awareness of their own oppressive norms "she has adopted as a part of herself conception and motivational structure" (82). This empathetic dialogue, says Atkins, provides tools that do not dismiss her capacity for agency, "striking a crucial balance", according to the authors (82).

The reality, say the authors, oppressive socialization may have been so detrimental to an individual that this critical examination may not be as successful or possible as with others. While this may be the case for some, however, a "patient's choice may […] be less than completely competent and voluntary, […] even this is an improvement on the standard model of informed consent, which has no way of even identifying this problem with oppressive socialization" (83). Say Sisti and Stramondo, "as patient-centered care drives new health-care payment models and policy regimes, it would seem appropriate to identify the weaknesses of the standard model of informed consent, understanding the role of oppressive socialization, and then encourage more holistic notions of autonomy to take root" (83).

c. Vulnerability and Consent: A summary and discussion on how to these factors should be considered going forward in the data collection and privacy of individuals

The ability to categorize rests in a paternalistic power — where, in order to control a population, power is used to break the population down into smaller manageable, and therefore

further weakened parts. “The process of classification itself is a demarcation of power, an organization of knowledge and life that frames the conditions of possibilities of those who are classified” (Cheney-Lippold, 196).

Overall, in reviewing Rogers *et al.* and Sisti and Stramondo, we can conclude that while we are all vulnerable to some degree, just by being human as “human life is conditioned by vulnerability” (12), it is important to differential between this inherent vulnerability and situational or dispositional vulnerabilities. The need for identifying different sources of vulnerability and different ways it is realized would work to ensure that morality responds appropriately, otherwise the responses run the risk of doing too little or too much, enlisting paternalistic protections. According to Rogers *et al.*, the duties need to be rooted in this understanding. Further, any analysis should understand vulnerability as the exposure and risks to serious harms to one’s vital needs. As asserted by Sisti and Stramondo, the goal of increasing or enabling the autonomy of those vulnerable populations and restoring their agency.

For Sisti and Stramondo, the concept of autonomy is focused on the ability of vulnerable persons to give consent and how oppressive socialization and norms have affected both individual’s understanding of their own shortcomings in the consent they give and society’s thresholds for individuals to give informed consent. Reviewing Sisti and Stramondo, competence is a slippery slope for informed consent, because either: 1) its presence assumes that an individual understands all of the contexts and implications they are agreeing to, regardless of the oppressive socializations they have been exposed to (aware of them or not); or, 2) its absence assumes the individual incapable of giving consent and their ability to do so is removed from them. This is coupled with the fact that, for example, shame is often a result of eroded self-worth and therefore an eroded competence, which further exacerbates and makes the individual

vulnerable. Similarly, as per Rogers *et al.*, the danger of even just identifying a population or individual as vulnerable as an instigating of paternalistic protections.

The policies designed for all individuals, in particular vulnerable populations, should work to protect them from further vulnerabilities, not cause them. Doing so would at the same time recognize the different forms of vulnerability and the needs and harms involved in them would address the sources and causes of vulnerability – and this would work to end the systemic issues in working with these populations.

Sisti and Stramondo encourage the engagement of the patient in helping them find the words and descriptions they need to express themselves, which would ultimately help them – at a minimum – identify their oppressive norms, resulting in an increased capacity for agency. Similarly, Rogers *et al.* insist that instead of seeing a population as vulnerable, as an identifying factor and a state of being, it is important to instead see their position as a result of their vulnerabilities. In doing so, it would be more likely to see the individual as more than the societal factors that count against them. In this case, their autonomy and agency would be better enabled if viewed as individuals who are affected by the vulnerabilities around them, and that the removal of these vulnerabilities would, and should, be the goal of policies designed and implemented for these populations. Further to this, as per Sisti and Stramondo, their ability to give consent should not be removed – removing their autonomy – but rather enabled with an awareness of the limitations, which would work to enable or increase their autonomy. In doing so, even a reduced level of competence and voluntariness in the ability to give consent is better than having the ability removed completely. As stated by Cheney-Lippold, “when identity is formed without our conscious interaction with others, we are never free to develop — nor do we know how to develop. What an algorithmic gender signifies is something largely illegible to us,

although it remains increasingly efficacious for those who are using our data to market, surveil, or control us" (219).

5. CURRENT LEGISLATION AND POLICIES ON PRIVACY AND DATA AND THEIR IMPLICATIONS

a. Canada – An introduction to PIPEDA and related documents

The Personal Information Protection and Electronic Documents Act (PIPEDA) received Royal Assent in April, 2000. The document governs the collection, use and disclosure of the personal information of Canadians by private-sector organizations. Its compliance is overseen by the Privacy Commissioner of Canada and the Act itself is the responsibility of the Minister of Innovation, Science and Economic Development. The Act was amended in 2015 to include items such as the disclosure of information on an individual, even "without their consent or knowledge, to an investigative body when there are reasonable grounds to believe that the information related to breach of an agreement or a contravention of the laws of Canada, a province or a foreign jurisdiction that has been, is being or is about to be committed" (Canada, Bill S-4). The then Privacy Commissioner found these permitted disclosures without consent, at the time and in the name of fraud prevention, to be problematic. The Commissioner highlighted the risk that this "may open the door to widespread disclosures and routine sharing of personal information among organizations based on a hypothetical risk of fraud" (Canada, Bill S-4). The Act also put in accountability safeguards, that include a need for justification by the investigators, a third-party reviewer and a compiled list of the eligible investigative bodies at all times (Canada, Bill S-4). To this, the Commissioner highlighted the lowered threshold required for investigation. The Commissioner called for, "disclosing organizations […] be required to report publicly on the number of disclosures being made and the types of organizations involved. Disclosing organizations should also be required to document the analysis undertaken in deciding to disclose information under this provision. These mechanisms would aid in holding organizations

accountable for disclosures that would otherwise be invisible" (Canada, Bill S-4). This lead to the Commissioner calling for a legal framework on the matter, citing its need for "clarity and guidance. […] Such a framework would provide Canadians with greater transparency about private sector disclosures of personal information to state agencies" (Canada, Bill S-4).

i) TRUST BUT VERIFY

In September 2018, the Office of the Privacy Commissioner, in its Annual Report to Parliament, revisited PIPEDA and declared the need for its revamping and re-visioning. Citing new advances in technology coupled with recent data scandals, the Commissioner declared, "these issues also underscore deficiencies in Canada's privacy laws that I and my predecessors have tried to draw attention to for years" (Canada, 2017-18 Annual Report to Parliament, 1). The Report highlighted how recent privacy laws have been "too permissive" for companies and that the "time for self-regulation is over" (Canada, 2017-18 Annual Report to Parliament, 1). This is not to say that environment for creativity and private company development is going to be disregarded in future legislation. Rather, according to the Report, the government needs to take a more active role as regulator for these companies to protect Canadians. "In other words, trust but verify" (Canada, 2017-18 Annual Report to Parliament, 2).

The Report cites the recent scandals around privacy (such as those aforementioned in this thesis involving Cambridge Analytica and Facebook) as clear indicators that such revisions are necessary. Accompanying the Report are several Guidelines, discussed later in this Chapter, that support the Act in a time of rapidly changing technologies and which were a result of consultations held by the Office of the Privacy Commissioner. The Report also cites a new "proactive vision for privacy protection" which has created two new program areas of promotion

(bringing federal departments and other organizations towards compliance with the laws on privacy) and compliance (addressing existing compliance issues).

In the OPC's Annual Report to Parliament on Consent (2016-17), it was concluded that, "consent remains central to personal autonomy and should continue to play a prominent role in privacy protection, where it can be meaningfully given with sufficient information" (Canada, 2017-18 Annual Report to Parliament, 11). The Report goes on to say that consent, "must be supported by other mechanisms if we are to effectively protect privacy, including independent regulators that inform citizens, guide industry, hold it accountable, and sanction inappropriate conduct" (Canada, 2017-18 Annual Report to Parliament, 11). Based on this, the Office created the previously mentioned Guidelines, specifically: the Guideline on Inappropriate Data Practices; the Guideline on Obtaining Meaningful Consent; and, the Draft Position on Online Reputation. These Guidelines have since come into effect and the Draft Position, as will be explained shortly, wades into debate around the Right to be Forgotten, a central piece of this thesis.

PIPEDA complaints accepted* by industry sector

Industry sector	Number	Proportion of all complaints accepted **
Accommodations	19	6%
Entertainment	5	2%
Financial	70	24%
Food and beverage	2	1%
Government	2	1%
Health	4	1%
Individual	1	0%
Insurance	21	7%
Internet	31	10%
Manufacturing	4	1%
Not-for-profit organizations	1	0%
Professionals	18	6%
Publishers	3	1%
Sales/retail	16	5%
Services	43	14%
Telecommunications	40	13%
Transportation	17	6%
Total	**297**	**100%**

* PIPEDA complaints accepted based on count of one for each series of complaints dealing with related issue; excluded complaints total six

** Figures may not sum to total due to rounding

Table 1: (Canada, 2017-18 Annual Report to Parliament, 72)

PIPEDA complaints accepted* by complaint type

Complaint type	Number	Proportion of all complaints accepted**
Access	86	29%
Consent	70	24%
Use and disclosure	62	21%
Safeguards	45	15%
Collection	15	5%
Retention	5	2%
Accuracy	5	2%
Openness	3	1%
Accountability	2	1%
Correction/notation	2	1%
Appropriate purposes	1	0%
Other	1	0%
Total	**297**	**100%**

* PIPEDA complaints accepted based on count of one for each series of complaints dealing with related issue; excluded complaints total six

** Figures may not sum to total due to rounding

(Table 2: Canada, 2017-18 Annual Report to Parliament, 73)

Privacy Act dispositions of access and privacy complaints* by institution

Respondent	Discontinued	No jurisdiction	Not well-founded	Resolved	Settled	Well-founded	Well-founded resolved	ER-resolved	Total
Administrative Tribunals Support Service of Canada								1	1
Agriculture and Agri-food Canada								1	1
Bank of Canada								1	1
Canada Border Services Agency	4		2		2		3	36	47
Canada Mortgage and Housing Corporation			1						1
Canada Post Corporation	4					1		15	20
Canada Revenue Agency	8		13	1		1		31	54
Canada School of Public Service					1			1	2
Canadian Air Transport Security Authority								1	1
Canadian Broadcasting Corporation								1	1
Canadian Food Inspection Agency							1	5	6
Canadian Human Rights Commission							1	2	3
Canadian Human Rights Tribunal	1								1
Canadian Museum of History	1								1
Canadian Radio-television and Telecommunications Commission	1								1
Canadian Security Intelligence Service	1		9		1			20	31
Communications Security Establishment Canada			1					3	4
Correctional Service Canada	6		9	2	6	6	5	56	90
Department of Justice Canada			3		1		2	6	12
Department of National Defence	4		5	1	4	1		19	34
Elections Canada								1	1
Employment and Social Development Canada	4		5		1	1	1	14	26
Environment and Climate Change Canada			1	4					5
Financial Transaction and Reports Analysis Centre of Canada			1					2	3
Fisheries and Oceans Canada			1			1		1	3
Global Affairs Canada								1	1
Health Canada			1					4	5
Immigration and Refugee Board of Canada	2		1					4	7

(Table 3a: Canada, 2017-18 Annual Report to Parliament, 79)

Respondent	Discontinued	No jurisdiction	Not well-founded	Resolved	Settled	Well-founded	Well-founded resolved	ER-resolved	Total
Immigration, Refugees and Citizenship Canada	1		4		3			20	**28**
Indigenous and Northern Affairs Canada								5	**5**
Innovation, Science and Economic Development Canada	2							1	**3**
Library and Archives Canada								6	**6**
Marine Atlantic Inc.								1	**1**
National Energy Board								1	**1**
National Research Council of Canada					1			2	**3**
Natural Resources Canada			1	1		1		2	**5**
Office of the Commissioner of Official Languages							2		**2**
Office of the Correctional Investigator								2	**2**
Office of the Information Commissioner of Canada	1					2		1	**4**
Office of the Public Sector Integrity Commissioner of Canada								1	**1**
Parks Canada Agency				1				2	**3**
Parole Board of Canada	1		1			1	1	8	**12**
Privy Council Office			1			1	1	2	**5**
Public Health Agency of Canada								1	**1**
Public Safety Canada			2					2	**4**
Public Service Commission of Canada				1				6	**7**
Public Services and Procurement Canada	1					2	1	8	**12**
Royal Canadian Mounted Police	6	1	4		3	10	3	48	**75**
Security Intelligence Review Committee								1	**1**
Service Canada			1					3	**4**
Shared Services Canada								3	**3**
Statistics Canada	1							9	**10**
Sustainable Development Technology Canada			1						**1**
Transport Canada	2		1	1			1	6	**11**
Treasury Board of Canada Secretariat								1	**1**
Veterans Affairs Canada		2	4	1		2		10	**19**
Veterans Review and Appeal Board								1	**1**
VIA Rail Canada								2	**2**
Western Economic Diversification Canada								1	**1**
Total	**51**	**3**	**73**	**13**	**23**	**30**	**22**	**382**	**597**

* PA complaints closed based on count of one for each series of complaints dealing with related issue; excluded complaints total 26

(Table 3b: Canada, 2017-18 Annual Report to Parliament, 80)

The Report goes further into providing statistics and examples of cases the Office of the Privacy Commissioner has reviewed and ruled on. As seen in Table 1 and 2, the Report outlines complaints received through the Office of the Privacy Commissioner both by Sector and by complaint type. In the complaints by Sector [Table 1], the financial sector tops the list of numbers of complaints at 24%, followed by services at 14% and telecommunications at 13%. In the list of complaint type [Table 2], access at 29% tops the list (where, "the institution/organization is alleged to have denied one or more individuals access to their personal information as requested through a formal access request"), followed by consent at 24%, (where, "under PIPEDA, an organization has collected, used or disclosed personal information without valid consent, or has made the provisions of a good or service conditional on individuals consenting to an unreasonable collection, use, or disclosure"), followed by use and disclosure at 21% (where, "the institution/organization is alleged to have used or disclosed personal information without the consent of the individual or outside permissible uses and disclosures allowed in legislation") (Canada, 2017-18 Annual Report to Parliament, Appendix 70-3). These numbers are interesting in that they point to the importance of consent, access as related to privacy and discussed in this thesis and point to the need for a re-examination of the legislation, focussing on the combined interests around access to data, the consent to collect data and the use and disclosure of that data. Of further interest in the tables of the report [Tables 3a&b] and for this thesis, the breakdown of PIPEDA complaints by federal department shows that the top three are Correctional Services Canada, with 90% of complaints, and the RCMP, which accounts for 75% of complaints, followed by the Canada Revenue Agency with 54% of complaints (Canada, 2017-18 Annual Report to Parliament, Appendix p. 79-80). These statistics are important to reflect on given the issues surrounding the case study of this thesis and the need to focus on

persons with increased vulnerabilities in reviewing issues around privacy, autonomy and authority.

The Report also draws specific reference to one case in particular that draws correlations to the Right to be Forgotten in *Google Inc. v. Equustek Solutions Inc.* (2017 SCC 34). The case involved a company (Equustek Solutions Inc.) which wanted to have Google remove all references to another company called Datalink, which Equustek, "alleged that, in addition to trademark infringement though re-labelling ESI's product and passing it off as its own, Datalink unlawfully obtained confidential information and trade secrets belonging to ESI. Datalink then used this information to create and sell a competing product, the "GW1000"", according to a review of the case by Gowlings WLG, an international and renowned law firm ("The Google Inc. v. Equustek Solutions Inc. Decision"). "In this case, the court ruled that it was possible for a Canadian court to grant a worldwide interlocutory injunction against a search engine in order to have it delist websites. While the case had to do with trade litigation, it has implications for privacy and may have ramification for the "right to be forgotten" debate" (Canada, 2017-18 Annual Report to Parliament, 14). The connection with the RtbF debate was that the court case led to the precedent that a search engine can be ordered to "remove (de-index) information from its search results" (Canada, 2017-18 Annual Report to Parliament, 14). The impact, according to the Office of the Privacy Commissioner, this case had is on the direction that privacy legislation has and may further take towards supporting a RtbF in the future.[6]

[6] This is debatable, as laid out by Michael Geist in "Does Privacy Law Apply to Google Search?", where on the article Geist states, "but the court is not being asked whether the current law includes a right-to-be-forgotten. Instead, the very application of Canadian privacy law to Google search is at stake." Concludes Geist, "The analysis suggests that the Privacy Commissioner's reference is no slam dunk. Indeed, there are strong arguments that PIPEDA does not apply to the search indexing and display. The right to be forgotten is problematic for several reasons, but the issue – along with the limited scope of PIPEDA – would be better addressed as part of a long overdue review and update to Canada's privacy laws."

Of overall interest is that fact that the Report appears to cite the GDPR (the EU's General Data Protection Regulation, introduced later in this thesis) as the goal of House of Commons committee on Ethics (ETHI) and that, "the committee went beyond our recommendations for reforms, effectively calling for changes that would more closely align PIPEDA with the European Union's GDPR" (Canada, 2017-18 Annual Report to Parliament, 14). The Report cites the committees call for "privacy by design"[7] to help better enforce data rights for individuals. The Report also cites the Office of the Privacy Commissioner's call, to ETHI, "for maintaining Canada's "adequacy status" with the European Union" (Canada, 2017-18 Annual Report to Parliament, 14). However, clarifies the Report, "since 2001, data has been allowed to flow freely from the European Union to Canada. With the GDPR – the new European data protection instrument now in force – decisions about the adequacy of a country's privacy laws, in terms of whether they afford European citizens protections equal to those of Europe, will be reviewed every four years. The committee was receptive to these concerns and called on the government to take appropriate action to ensure that the seamless transfer of data between Canada and the European Union can continue." (Canada, 2017-18 Annual Report to Parliament, 14-5).

ii) GUIDELINE ON INAPPROPRIATE DATA PRACTICES (PIPEDA Subsection 5(3))

The "Guideline on Inappropriate Data Practices" was released as part of the Office of the Privacy Commissioner's efforts to improve areas around the issue of consent under PIPEDA's reach. The Guideline came into effect on July 1, 2018. It introduces "No-go zones" (Canada,

[7] In this context, "privacy by design" refers to Dr. Ann Cavoukian's, the Ontario Information and Privacy Commissioner in 2007, argument for the need for developers to "build privacy concerns right into their work" (Canada, Office of the Privacy Commissioner of. *Privacy by Design*. 8 Oct. 2007, https://www.priv.gc.ca/en/blog/20071008/).

Guideline on Inappropriate Data Practices) as a boundary around what kind of information an organization can collect, use and disclose personal information, separating data which is "legitimate" and required by law from data which is off-limits. The document provides the principles around these "No-go zones" for organizations to operate around and requires them "to engage in a "balancing of interests" between the individual and the organization concerned" as ""viewed through the eyes of a reasonable person." (Canada, Guidance on Inappropriate Data Practices).

The Guideline outlines how consent from an individual is not enough for an organization to demonstrate its compliance with PIPEDA, saying consent is "necessary but not sufficient" (Canada, Guideline on Inappropriate Data Practices). The Guideline requires context when reviewing the data practices of an organization, examining the circumstances around which the data was collected, used or disclosed. The document sets out the following factors, based in legal precedence, for evaluating the compliance of an organization:

- the degree of sensitivity of the personal information at issue;
- whether the organization's purpose represents a legitimate need / *bona fide* business interest;
- whether the collection, use and disclosure would be effective in meeting the organization's need;
- whether there are less invasive means of achieving the same ends at comparable cost and with comparable benefits;
- and, whether the loss of privacy is proportional to the benefits (Canada, Guideline on Inappropriate Data Practices).

Beyond an organization's requirements for compliance, the Guideline also lays out the specifics around its "No-go zones", the purposes for which the collection, use or disclosure of personal information "would generally be considered "inappropriate" by a reasonable person" (Canada, Guideline on Inappropriate Data Practices). They are listed as:

1) Collection, use or disclosure that is otherwise unlawful;

2) Profiling or categorization that leads to unfair, unethical or discriminatory treatment contrary to human rights law;

3) Collection, use or disclosure for purposes that are known or likely to cause significant harm to the individual;

4) Publishing personal information with the intended purpose of charging individual for its removal;

5) Requiring passwords to social media accounts for the purpose of employee screening; and,

6) Surveillance by an organization through audio or video functionality of the individual's own device (Canada, Guideline on Inappropriate Data Practices).

According to the Guideline and in summary, "an appropriate purpose judged from the standpoint of a reasonable person is a flexible concept that requires time, careful reflection and practical experience to define. In practice, the test for appropriateness will require a contextual analysis but we find it useful—for transparency to both individual and organizations—to provide examples of our expectations, such as those listed above. It is our intention to periodically revisit

and update the above list of "No-Go zones" as warranted" (Canada, Guideline on Inappropriate Data Practices).

The highlight of the Guideline is that it requires compliance of an organization when consent has been received from an individual. As will be discussed later in this Chapter, this is an area where the RtE and the RtbF focus on in their work to protect an individual's privacy and autonomy. Most interesting from the "No-Go Zones" listed above are items 2 ("Profiling or categorization that leads to unfair, unethical or discriminatory treatment contrary to human rights law") and 3 ("Collection, use or disclosure for purposes that are known or likely to cause significant harm to the individual"). It can be argued that, by applying a feminist bioethical framework on vulnerability, discriminatory treatment based on profiling is increased as the vulnerabilities an individual is exposed to are increased. Further, "significant harm" can be a subjective term in this way, as harm can be a weighted term. It could be determined, even by a "reasonable person" that the harm an individual exposed to is either insignificant or relative to the goals of the organization or the potential benefits and outcomes projected. In terms of the case study of this thesis, the harm to the individual in crisis could be considered — as perceived by the compliance reviewer — as relative and therefore acceptable to the benefits the program provides, both for the safety of the individual and for the support of front-line workers. The RtE and the RtbF, as will be discussed, go further than this in their protections of the privacy of individuals.

iii) GUIDELINE FOR OBTAINING MEANINGFUL CONSENT

The "Guideline for Obtaining Meaningful Consent", which came into effect on January 1st, 2019, describes a strain being put on privacy laws by evolving technologies. As such, the Office of the Privacy Commissioner is revisiting the meaning and application of consent as it

applies to these technologies. In the Guideline, the Office of the Privacy Commissioner states that, “advances in technology and the use of lengthy, legalistic privacy policies have too often served to make the control – and personal autonomy – that should be enabled by consent nothing more than illusory” (Canada, Guidelines for Obtaining Meaningful Consent). The Guideline calls on companies to create new ways for developing consent processes that are in line with the following principles:

1) Emphasizing key elements;

2) Allow individuals to control the level of detail they get and when;

3) Providing individuals with clear options to say ‘yes’ or ‘no’;

4) Be innovative and creative;

5) Consider the consumer’s perspective;

6) Make consent a dynamic and ongoing process; and,

7) Be accountable: Stand ready to demonstrate compliance (Canada, Guidelines for Obtaining Meaningful Consent).

The Guideline also cites the importance and the difference between forms of consent, express and implied, as a critical point that needs to be taken into consideration by companies and organizations. Express consent should be required when information is sensitive and its “collection, use or disclosure is outside of the reasonable expectations of the individual” that may “create a meaningful residual risk of significant harm” (Canada, Guidelines for Obtaining Meaningful Consent). The report also highlights the fact that “even non-sensitive information can become sensitive depending on the circumstances” (Canada, Guidelines for Obtaining

Meaningful Consent). “Reasonable expectation” links to the understanding of the use of their data by the individual. For example, an individual might “reasonably expect that information could be disclosed to a third party with a legal entitlement to it; however, an individual would not reasonably expect disclosure to individuals who are merely curious or seek the information for nefarious purposes” (Canada, Guidelines for Obtaining Meaningful Consent).

The “Risk of Harm” factor includes harm to the individual not just to property, for example, but also to reputation. If there was the potential risk to reputation for an individual, the consent would need to be express, not implied. The age of the individual is also brought into play, where children are unable to give consent but it must be “obtained from their parents or guardians” (Canada, Guidelines for Obtaining Meaningful Consent). The Guideline also highlights that the use, collection and disclosure of personal information must be appropriate, defined and limited even when consent is received, meaning that consent is not a “free pass” to use personal information. Further, the Guideline states the importance for the ability of an individual to withdraw consent, meaning that upon request further information should not be collected and in fact previously collected information should be deleted. The problematic aspect to this is highlighted, wherein some, “laws may require that information be retained” for example in the financial sector wherein credit applications need to be retained for five years (Canada, Guidelines for Obtaining Meaningful Consent).

iv) OPC POSITION ON ONLINE REPUTATION

In the draft paper titled, “Online Reputation - What Are They Saying about Me?” by the Office of the Privacy Commissioner of Canada and published in January 2018, the office names

"Reputation and Privacy" as one of its strategic privacy priorities for 2015-2020 in their goal of updating the PIPEDA. They set their goal as being that they "will have helped to create an environment where individuals may use the Internet to explore their interests and develop as persons without fear that their digital trace will lead to unfair treatment" (Canada, Online Reputation). This article is important because it outlines what Canada is doing and what it says it wants to do regarding the protection of Canadians' data privacy.

"[T]he Office of the Privacy Commissioner's draft position highlights existing protections in Canada's federal private sector privacy law, identifies potential legislative changes and proposes other solutions for consideration" (Canada, Online Reputation). The draft highlights three areas of focus in protecting the online reputation of individuals: 1) de-indexing; 2) source removal; and, 3) privacy education. De-indexing is the process by which a webpage, image or other online resource is removed from search engine results when an individual's name is entered as the search term, as was seen in the *Equustek v. Google Inc.* case mentioned earlier in this Chapter. This is covered by PIPEDA and "includes allowing individuals to challenge the accuracy, completeness and currency (the extent to which the information is up-to-date) of results returned for searches on their name. The OPC states that such challenges should be evaluated on a case-by-case basis. Second in the draft paper is "source takedown". Simply put, it is the removal of content from the internet. As updated PIPEDA would provides individuals the right to withdraw consent and requires that personal information that is no longer needed be destroyed, erased or made anonymous. The report states that, "taken together, this implies that individuals should have the ability to remove information that they have posted online" (Canada, Online Reputation). The report identifies differences about information posted by others about them, but states that people can challenge the accuracy and completeness of the information

through a formal complaint with the Office of the Privacy Commissioner. Other facets of the approach include: an absolute right for youth to remove any content from the internet posted by them (at any time) or their parents/guardians (when they have reached the age of majority)); the need to focus on privacy where vulnerable groups are concerned; and the development of an industry-wide code of practice be developed for takedown policies, privacy defaults and procedures.

Focusing on vulnerable populations, the Paper draws direct connections with this thesis in that it addresses the "disproportionate impact on vulnerable individuals", the factors surrounding their ability to consent to or contest their data being collected or used and how these combine to create greater potential issues for their reputations:

> [W]e note that the OPC has clearly indicated its intention to take a more proactive role in ensuring compliance with PIPEDA, to address broader, systemic privacy risks to Canadians. This will be of particular importance in the realm of Online Reputation, given the number of individuals who may be impacted by a single non-compliant organization, and the significant impacts on individuals that can result from non-compliance, particularly the disproportionate impact on vulnerable individuals, who may opt not to file complaints to our Office in order not to bring further attention to themselves, or open themselves to retribution. Each of these factors point to the benefits of curbing privacy risks up front, before problems occur, or in high-risk areas where problems may not be outwardly detectable" (Canada, Online Reputation).

With this points raised, the Paper makes a direct reference with the RtbF, introduced for further analysis in this thesis. It states that, "while, in combination, the abilities to request de-

indexing and/or source takedown of information in certain circumstances are similar to the "Right to Erasure (RtbF)" in the EU's General Data Protection Regulation (GDPR), this paper does not import a European framework into Canada" (Canada, Online Reputation). This is an interesting perspective for the Office of the Privacy Commissioner to take, and one that expands beyond the purview of this thesis. While they do have the same goals, and the paper does take direct action on how or if it would work to incorporate the GDPR's RtbF beyond its supports for de-indexing and source takedown. The Office of the Privacy Commissioner ultimately leaves it for a decision to be reviewed and decided on by elected officials.

Ultimately, the Office of the Privacy Commissioner's draft report concludes that Canadians have an existing right under PIPEDA to ask search engines to de-index web pages, and to ask websites to remove or amend content that contains inaccurate, incomplete or outdated information. The proposal draws parallels with the RtbF in the European Union. However, it stops short of expanding to address not just online posting but also the data points collected and hidden behind the scenes, like that of the RtbF in the GDPR (General Data Protection Regulation) as will be reviewed later in the paper.

b. Europe – Overview and introduction of GDPR, the RtE and the RtbF

Currently, the GDPR (General Data Protection Regulation) is an EU legislation that imposes on data controllers requirements around the data and use its collects from its subjects. 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. Debate on the legislation brings up questions about whether it implies a RtE 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" ("Recital 71 – Profiling").

The question central to the debate around the RtbF and the GDPR is whether current and new regulations around an individual's right to access and protect their data then creates a RtE for any decisions made through automated decision-making systems for that individual. For the purposes of this thesis, I will not debate whether a RtE exists in the GDPR — rather, I will argue that a RtE should exist for what can still be called 'data subjects', in particular those subjects whose data is collected due to their vulnerability, social access and status. At this point, I will introduce the rational behind the Rights to Explanation (RtE) and to be Forgotten (RtbF).

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, catalog 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" (32). Among the many ethical considerations surrounding privacy and an individual's data use and storage, the RtE and the RtbF are being debated and contested in the EU, which is currently examining rights surrounding its citizens' data in the aforementioned GDPR. These Rights

underscore the importance of ethical challenges in the use of algorithms and data for automated decision-making systems and are important to examine as they become the intersection between the importance of data collection, its use for both positive and negative societal gains and the where we need to draw the line at where a persons privacy ends and where (or whether) the benefits that data can provide begin, outweighing the effects of its invasion.

i) The Right to Explanation

In Margot Kaminski's article, "The Right to Explanation, Explained", the author introduces the RtE in the context of the debate surrounding its presence in the GDPR. The article clearly outlines the portions of the GDPR that permits the RtE and what they mean in the debate around algorithmic governance and accountability. The article also further delves into the interpretation of the regulation and the debate surrounding whether the RtE actually exists within it. For the purposes of this thesis, I will be focussing on the definitions of what a RtE is and how it provides transparency in the face of the opaqueness of black boxes, becoming an ethical necessity, particularly for those citizens in vulnerable populations. "Transparency is a basic principle of the GDPR […] data protection regimes are grounded on fairness, and transparency and fairness are linked ideals; we often use transparency as an element of accountability, to establish that systems are fair" (Kaminski, 17-8).

The General Data Protection Regulation (GDPR), which came into effect on May 25th, 2018, "contains a significant set of rules on algorithmic accountability, imposing transparency, process and oversight on the use of computer algorithms to make significant decisions about human beings" (Kaminski, 2). Kaminski highlights 4 Articles in the GDPR that deal explicitly with the RtE: Article 13 ("establishes notification rights/requirements when information about individuals is collected directly from individuals"); Article 14 ("establishes notification

rights/requirements when information is collected from third parties"; Article 15 ("creates an individual right of access to information held by a company that can be invoked "at reasonable intervals"); and Article 22 (which "states that individuals have the right not to be subject to a decision based solely on automated processing") (4). The first three Articles, according to Kaminski, require disclosure of "the existence of automated decision-making, including profiling", while Articles 13 and 14 create a requirement for companies to inform individuals when their data is obtained and Article 15 "creates an access right at almost any time" (7). Article 22, according to Kaminski, has proven more problematic in its interpretation. The question surrounds whether an individual has the right to object to the use of ADMSs or whether they have the right to know one is being used and how it may effect any potential decisions or outcomes involving them. The right only applies, according to the Article, when a decision is "'based-solely' on algorithmic decision-making [a]nd it applies only when the decision produces 'legal effects' or 'similarly significant' effects on the individual" (Kaminski, 5). The Article also has three exceptions to this right: when the automated decision is "'necessary… for a contract'"; when the [EU] Member State has passed a law "creating an exception"; and, when an individual has, "explicitly consented to a decision". Other portions of the Article highlight the "right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (Kaminski, 5). Kaminski articulates that, "this explicitly creates a version of algorithmic due process: a right to be heard" (6). As will be seen later in this Chapter in the discussion on the RtbF, individuals will have further opportunities to overcome these exceptions.

Article 22 clarifies the duty of companies to not solely use automated decision-making, which will require human oversight to be, "carried out by someone who has the authority and competence to change the decision" and they "must additionally have access to information

beyond just the algorithm's outputs", which will "thus have the effect of requiring companies to think about how they structure their "human in the loop" of algorithmic decision-making" (Kaminski, 11).

The guidelines provide both a framework for determining what constitutes a significant effect, and a list of examples: decisions that affect financial circumstances, decisions that affect access to health services, decisions that deny employment or put someone "at a serious disadvantage" (Kaminski, 11), and decisions that affect access to education. The guidelines also explain that some behavioral advertising will be covered. Particularly intrusive advertising targeted at particularly vulnerable data subjects in particularly manipulative ways will trigger Article 22. Differential pricing—showing people different prices based on personal profiles—could also trigger Article 22 if "prohibitively high prices effectively bar someone from certain goods or services" (Kaminski, 11-2). "Thus Article 22's algorithmic accountability provisions will reach both at least some behavioral advertising and some differential pricing tactics. This is broader coverage than some scholars predicted" (Kaminsky, 12).

Kaminski also reviews how the GDPR accounts for the concerns surrounding 'trade secrets', a common reasoning deployed by companies to avoid having to reveal the algorithms behind the resulting decision. "[S]everal scholars feared that in practice, companies could avoid the GDPR's transparency requirements by citing a need for corporate secrecy […] while there is "some protection" against having to reveal trade secrets, companies "cannot rely on the protection of their trade secrets as an excuse to deny access or refuse to provide information"" (Kaminski, 12). Clarifies Kaminski, "While this does not eliminate the trade secrets exception […] it does at least urge data protection authorities to watch for the use of overly broad trade secrets claims. Further, regarding the right to human review granted through Article 22,

companies cannot rely on the argument that the quantity of information processed is sufficient reason for automated decision-making systems, and "must show that there is no other effective and less privacy-intrusive way to accomplish he same goal"" (Kaminski, 12). According to Kaminski, the guidelines also work to address profiling, ensuring that the individuals are informed on their potential exposure to the risks associated. "The guidelines similarly constrain the explicit consent exception and turn it into an information-driving tool. They explain that individuals must be provided enough information about the use and consequences of profiling to ensure that any consent "represents an informed choice" (Kaminski, 12). The guidelines do not provide additional information about "explicit consent," except to note that while explicit consent is it is not defined in the GDPR, a "high level of individual control over personal data is… deemed appropriate" (Kaminski, 12). Strengthening this further is the claim that "suitable safeguards… should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. […] The guidelines counsel that there is a need for this form of transparency because an individual can challenge a particular decision or express her view only if she actually understands "how it has been made and on what basis." In other words, an individual has a RtE of an individual decision because that explanation is necessary for her to invoke the other rights—contestation, expression of her view—that are explicitly enumerated in the text of the GDPR" (Kaminski, 13).

The Article goes further to interpret "suitable safeguards", including "systemic accountability measures such as auditing and ethical review boards", providing systemic accountability [involving] internal auditing, quality assurance checks, third-party auditing and more" (Kaminski, 14). It charges companies "with preventing discrimination on the basis of race,

ethnic origin, political opinion, religion and more. The guidelines envision ongoing testing and feedback into an algorithmic decision-making system to prevent errors, inaccuracies, and discrimination on the basis of sensitive data" (Kaminski, 14-5).

According to Kaminski, "the GDPR is, in large part, a collaborative governance regime. The text is full of broad standards, to be given specific substance over time through ongoing dialogue between regulators and companies, backed eventually by courts" (9). The regulation creates an "opportunity to be heard" for all those whose data is collected and used to make decisions about them, their present or their futures. As was seen in Chapter 4 and the discussion around vulnerable populations, this is a critical and necessary step. Asserts Kaminski, "the "who" and "why" of transparency in the GDPR dictates the what, when, and how. Individual transparency provisions, the guidelines make clear, are intended to empower individuals to invoke their rights under the GDPR. Thus while individuals need not be provided with source code, they need to be given far more than a one-sentence overview of how an algorithmic decision-making system works. They need to be given enough information to be able to understand what they are agreeing to (if a company is relying on the explicit consent exception); to contest a decision; and to find and correct erroneous information, including inferences" (22).

Overall, in reviewing Kaminski's article within the parameters of this thesis, the RtE as outlined in the GDPR provides multiple outlets for the protection of an individual's information and privacy, resulting in the ability to ascertain how any data collected on them may be or may have been used negatively against them, both at the time of collection as well as after the fact, and how this impacts their autonomy. These outlets include: transparency in the eye of profiling; the act of holding a company responsible for providing a human reviewer; the right of an individual to be "heard"; the control against the argument companies hold for trade-secrets; the

necessity of consent as an informed choice; the building in of suitable safeguards, including systemic accountability measures and auditing, and; the resulting "collaborative governance regime" (Kaminski, 9) of the GDPR. All of these are aspects of the RtE in the GDPR that are critical for individuals in the protection of their information and privacy. Further, as was seen in the vulnerable framework portion of this thesis, these are critical points in addressing the needs of vulnerable populations.

According to Kaminski and her review of the RtE, communication to individuals about algorithmic decision-making must thus be simultaneously understandable (or "legible"), meaningful, and actionable. It must be understandable to individuals, rather than delivered in complex jargon or as an information flood. But it must also be meaningful in depth; the guidelines note that "[c]omplexity is no excuse for failing to provide information" (Kaminski, 21). And it must provide enough information that an individual can act on it—to contest a decision, or to correct inaccuracies or request erasure. Thus there is a clear relationship between the other rights the GDPR establishes—contestation, correction, erasure—and the kind of individualized transparency it requires. This suggests something interesting about transparency: that its substance is often determined by the substance of other underlying legal rights. If a person has a right of correction, that person needs to be able to see the errors. If they have a right against discrimination, they need to be able to see what factors are used in a decision. Otherwise, information asymmetries render underlying rights effectively void. The guidelines list examples of what kinds of information should be provided to individuals, and how it should be provided. Individuals should be told both the categories of data used in an algorithmic decision-making. process, and an explanation of why these categories are considered relevant. They should be told the "factors taken into account for the decision-making process, and… their respective 'weight'

in an aggregate level" (Kaminski, 22). They should be told how a profile used in algorithmic decision-making is built, "including any statistics used in the analysis" (Kaminski, 22) and the sources of the data in the profile. Individuals should be provided an explanation of why a profile is relevant to the decision-making process, and how it is used for a decision (Kaminski, 22).

ii) The Right to be Forgotten

In Bartolini and Siry's article, "The Right to be Forgotten in the light of the consent of the data subject", the authors discuss the evolution of the RtbF as it is now "statutorily proposed" in the GDPR and what it means for the concept of consent. The authors give brief history of the GDPR as an expanded interpretation of its previous incarnation as the DPD (Data Protection Directive). The expansion was necessitated by recent rapid technological advancement and their implications on data collection and use. The RtbF, or at least its sentiment, is believed to have existed in the DPD, though this is a debatable topic, as the authors demonstrate. The DPD introduced provisions such as the "right to rectification" and "the right to object", which stemmed from pre-existing EU data protection principles. The RtbF, however, is not a natural extension of these principles as is believed and this leads the authors to discuss the RtbF's impact on consent, which will be the focus of the discussion for this thesis.

There is a conundrum with the RtbF, highlighted by Bartolini and Siry, that adds a level of problems with both its interpretation and its impact. The authors cite that fact that the "processing of personal data requires that the data subject (DS) agrees by giving his or her informed consent" (219). In addition to this, the RtbF requires the erasure of personal data upon request of the data subject. This circular process implies that consent can be granted but then taken back or revoked, "on one side, giving one's consent is the door that opens up the lawfulness of the processing of personal data; on the other side, the willingness to be forgotten

(in the term of the GDPR) is the lock that makes further processing unlawful" (Bartolini and Siry, 219). A withdrawal of consent in this case removes the consent to process or collect an individual's data. However, the DPD, upon which the assumption of the RtbF is granted, is unclear as to whether or not consent at one point can be revoked at another. This is problematic as there is not a clear explanation of the process or limits surrounding the withdrawal of consent. While the DPD does allow for the right to object, say the authors, this arguably does not imply a right to withdraw consent (219). With this in mind, the article looks into the concepts of consent and the right to object and compares them to the RtbF.

The DPD of 1995 demonstrated an evolution of the relationship between personal data and its protection. Leading up to the DPD, the protection of personal data had two pieces to its genesis, one in the EU and one in the US. In the US its origins were part of a larger discussion on privacy, while in the EU its origins were separate from privacy and were instead about an individual's rights, founded in the belief that the processing of an individual's data was an extension of their rights. However, in that simpler time of data collection and processing, the volume of what it is today poses challenges to these concepts, even compared to the 1990s and the DPD.

The authors point to the problematic concept of consent as defined and applied by the DPD as the root of the problem of the RtbF today in the GDPR (220). Consent of a data subject was not as finite as it appears to be. The DPD allowed for the user's intent of the use of the data to factor into the processing of an individual's data. The data subject's intent is "meaningful" under fewer components of the DPD than that of the data controller. This was controlled for by requiring consent by the data subject before their data could be processed. The problem here: "the matter appears to be of mostly theoretical relevance. Especially in the light of the data

protection legislation, the relationship with defects of consent is independent of the contractual or non-contractual nature of the consent" (Bartolini and Siry, 221).

According to Bartolini and Siry, with the intent of the data user/controller factoring within the weight of a data subject's granting of consent, the stage was set for a debate around the rights of the data subject and the abilities of the data user/controller. The implications led to a lack of clarification around the interpretation and the application of the Protective. An example is that, "the DPD requires that the DS "must be given accurate and full information," and that the consent to data processing be given in a free and informed way. Article 10 contains provisions concerning information that must be given to the DS, applicable in all cases of collection of personal data, regardless of the specific purpose of the processing. Finally, consent is defined as a freely given and informed indication about the agreement to being processed. Together, these provisions imply that the contractual or non-contractual nature is not particularly relevant: in any case where the consent is given under a defective situation, the processing is unlawful" (Bartolini and Siry, 221)."

The conflicting interests, wherein granting consent in turn implies an "acquiescence of a right in favour of the other party's interests" (Bartolini and Siry, 221), is problematic. Say the authors, "consent as surrender is inappropriate in the scope of the protection of personal data" (222). "The DS does not simply abandon the right over his or her personal data by giving his or her consent. Consent has a more procedural function as far as data protection is concerned. By giving consent, the DS maintains a degree of control over the processing, as well as remedies in case of unlawful processing. The DS becomes actively involved in a dynamic relationship with the data controller to ensure that the processing is lawful, within the limits of consent, and fair" (Bartolini and Siry, 222). "In other words," say the authors, "the DS would not be entitled to

generically give up his or her control over any future processing and personal data by the data controller" (222).

Another aspect around consent raised by the authors is around the time or juncture at which consent should or can be granted by the data subject. "In the absence of any provisions, unless the DS expressly requests that data processing is allowed only for the future, it would be logical to assume that, once consent is given, it pertains to all processing carried out within the alleged purposes, regardless of whether it was previously illegitimate" (Bartolini and Siry, 222). This is pertinent as data is collected just by visiting a website, eg. in the form of cookies and called "up-front" processing, often without requesting or receiving consent by the user, which the authors say is considered unlawful by the DPD. When actual consent is received later in the 'relationship', new data points are merged with previously generated ones and the data subject has inadvertently given consent to data collected prior to them giving consent.

The authors also connect consent with a data subject's autonomy, wherein the original intention of consent would work to support an individual's autonomy — but today that is an "inefficient" (223) reality. Ensuring an individual's autonomy, say the authors, has many approaches, ranging from the protection of rights to the obligations of the data processor. Additionally, the authors raise an interesting point, "it has also been noted that acquiring the "informed consent" of the DS is no longer a viable solution to protect his or her personal data" (223).

So, what does this mean for the right of an individual to withdraw consent, which is in effect the RtbF, and the many factors surrounding it? Once a data subject has given consent that has "been given freely and based on fair and complete information", the Right to Object, which the "DPD requires that the data subject should have the right 'to object to processing in certain

circumstances'" applies" (Bartolini and Siry, 223). There are clarifications around the ability to apply this right – for example, the use of the data can determine whether an individual can apply this right, like in the cases where the data is used in the public interest, the right of the data subject are limited. However, on the whole, the DPD "does indeed grant the DS the right to object, but in a limited scope which can be further narrowed by national legislation" (Bartolini and Siry, 224).

Overall, the Right to Object can be imposed if it fits the following criteria:

> •he or she has a legitimate interest which outweighs those of the data controller, or the processing can potentially cause damage or distress, or, more generically, there are compelling and legitimate reasons to object. In this case, the DS may be required to provide a justification for the objection to be valid;
>
> • the processing is carried out for the purpose of direct marketing, or in some cases for market research. The DS can object to such processing at any time and without justification;
>
> • Member States can grant the right to object in situations not taken into account by the DPD, but this rarely occurs; in addition to this, the DPD allows the DS the right to rectification or erasure of incomplete or inaccurate data; however, this is different in that the data processor is not prohibited from further data processing even if the DS can prove this. (Bartolini and Siry, 224)

This, state the authors, shows that "it does not appear that the RtbF (as it is defined in the GDPR, thus allowing the erasure of the data and the propagation of the erasure request) can be inferred on the basis of a general application of the right to object. Various differences emerge

between the two, concerning both the prerequisites and the effects" (Bartolini and Siry, 225). These prerequisites (for the right to object) are objective and subjective in their requirements, as to the purpose of the data processing and the legitimate grounds (eg. proof of damage or distress). The difference here is that, "the right to object simply states that the processing 'may no longer involve those data' whereas the RtbF entitles the DS to obtain the erasure of the data and the propagation of the request" (Bartolini and Siry, 225).

Overall, according to Bartolini and Siry "the DPD does not explicitly grant the DS the right to withdraw consent. However, this does not mean that this right does not exist in the Directive. Member States can also implement it, since the spirit of the DPD allows them flexibility in raising the level of protection of the DS" (Bartolini and Siry, 225). Assert the authors, the implicit nature of the right to withdraw consent is problematic and leaves its interpretation and enaction up to the individual Member states (225).

This in turn returns to the question of consent as a whole. If an individual's consent is a condition of that individual's data processing, then how does this apply to its withdrawal? In this case, "once the consent is withdrawn, any further collection and storage of personal data is prohibited. The data controller is not allowed to retrieve any additional personal data about the DS, or make any use of those already acquired. Undoubtedly, the overall spirit of the DPD implies that no further data can be collected once the consent is withdrawn" (Bartolini and Siry, 227). The problem arises with data that is already collected with consent. Does this mean it must be deleted or simply retained without being processed further? Prior to the GDPR, this question was not directly answered, though the origins of what is now seen as a problem were evident in the DPD by virtue of the definition of "storage". Bartolini and Siry explain that the question rests on the concept of static or dynamic activity/storage, where static is the act of maintaining data

somewhere and dynamic is the activity that takes, moves, places and stores the data, both with or without further processing. According to the authors, static storage would be more consistent with the idea of consent withdrawal, where the static act does not include any portion of a processing action (227). Say the authors, "once the data have been stored, no operation occurs in simply maintaining them statically" (227).

Problematic with even the static definition is that it poses technical challenges to overcome that make it nearly impossible given for data storage, companies, etc. actually function and operate in the ways they do now and envision themselves to do in the future. The transfer of data is a constant in even something as routine as backing up files or migrating stored data between architecture for improvements. This clarification would need to be accommodated through approved "boundaries of legitimate processing" (227) in order to occur beyond a data subject's consent. However, even if this was done, then this is not a true RtbF, according to the authors. "The 'Right to be Forgotten and to erasure', as its title implies, requires the controller to erase all data pertaining to the DS, with full retroactive effect on data already collected" (Bartolini, 228). Ultimately, even the right to withdraw consent that was earlier given can be equated or used to support a RtbF, meaning that these are the points that need to be considered prior to a true RtbF being available to data subjects in the GDPR.

In the GDPR, there has been a "major overhaul" of data protections (Bartolini and Siry, 228). Of these, the RtbF is a controversial one. Originally, the RtbF was "the right to have any reference to data completely erased from publicly available communication services. However, this wording would have had unbearable consequences from a technical point of view therefore it has undergone major changes" (Bartolini and Siry, 228).

There are wordings that support the right to withdraw consent and a resulting erasure request of an individual's data without any justification. Included is wording that states, "'[the] data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal'" (Bartolini and Siry, 228). In this example, according to the authors, the data controller would not receive any legal penalty for any previous processing to this point.

While this sounds like a solution to earlier mentioned problems with the DPD, there are many problems highlighted by the authors which lead to ambiguities. Specifically, in Article 17 of the GDPR, such as:

> • the controller might not know or be able to contact all third parties;
>
> • third parties might have different grounds for the lawfulness of the data processing, so the erasure request might not be effective toward them even if it is for the original controller;
>
> • in the case of Internet bounces, it is still unclear who the third party controller responsible for the bounce actually is, whether the manager of the service or its users. Modern Internet has blurred the distinction between controllers and DSs, and this is a weak spot in data protection laws (Bartolini and Siry, 229-30).

With these ambiguities, and the recognition of weaknesses when it comes to third parties for example, the reality is that an individual's initial consent is an "irreversible condition" (Bartolini and Siry, 230) and full control will or can never be regained by the data subject. The RtbF is a critical attempt at restoring this control, by "granting the DS the power not only to decide who will be allowed to process his or her personal data (by giving consent), but also who

will no longer be allowed to process them (by requesting erasure). This is in line with the right to the protection of personal data granted by the ECHR [European Court of Human Rights]" (Bartolini and Siry, 230). The authors also point to other confounding aspects of the RtbF, such as it not being able to usurp freedoms of expression ("another fundamental right in the ECHR" (Bartolini and Siry, 230)), and that it cannot be used to impose censorship, a critical factor that is addressed in the GDPR.[8]

> "To sum it up: the reform proposal allows the DS to withdraw consent, at will and without conditions (unless the derogation described earlier applies). The controller must then erase all personal data pertaining to the DS, and forward the same request to data controllers that are known to be processing the data. Article 17 explicitly contains "RtbF" in the title. The question, at this point, is whether this is actually a right "to be forgotten", or it is not" (Bartolini and Siry, 230).

The scope of the RtbF is a new challenge address by the GDPR. While the term itself is not new (the authors cite its reference in several nation's legislations), it is new to the EU. Its enforcement could pose "major problems" (Bartolini and Siry, 230), like the third-party aspects mentioned earlier in this section. As it stands, according to Bartolini and Siry, the 'RtbF and to erasure', as its GDPR title implies, requires the controller to erase all data pertaining to the data subject exercising it, with full retroactive effect on data already collected. Ultimately, the

[8] Though outside the confines of this thesis, the ability to control ones image and reputation through the RtbF is proving to be problematic through debates on the topic. In erasing photos or data around oneself, an individual could remove, for example, evidence incriminating them in a crime or that may damage their professional reputation because of something they did earlier in their lives. This is perplexing in the issues is also listed as a problem for rights to free speech, among others. (See: Google Loses "right to Be Forgotten" Case. 13 Apr. 2018, BBC).
This adds a level of complexity for the RtbF but which, I believe, could be further strengthen with the examination through a feminist bioethical framework focusing on vulnerability. In doing so, this framework as applied to ADMSs could help to protect the rights of the individuals who were the victims of past crimes or actions, likely those more marginalized and with increased levels of vulnerability. However, the protection around any data that may apply to these cases would require further examination and does not fall within the areas discussed in this thesis.

removal of consent, including its retroactive removal, is required for a true RtbF, which, say Bartolini and Siry, comes from addressing this critical factor: “The problem becomes purely practical: knowing who the controllers processing the data are” (230). An obligation exists, within the GDPR, for the data collectors to know where the data may go from their hands, namely third parties. The tracking of data “bounces”, among other requirements in the GDPR, is required by the regulation through the creation and registering of ‘links’. In this way, the RtbF as anchored in the GDPR, “will simplify and embolden citizens’ right to control their image in the web” (Bartolini and Siry, 234).

c) Ethical Considerations – How the PIPEDA measures up to the Rights of Explanation and to be Forgotten

Search engines and websites are notorious collectors and users of individual data. 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 RtE and RtbF.

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. This would imply that the GDPR’s RtE and the RtbF would be of great interest to ETHI and the Office of the Privacy Commissioner. 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 that being adequate in the provisions it provides 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 RtE and RtbF if it hopes to address the ethical considerations for data collection and protect its citizens, particularly for vulnerable persons, as referenced in the Draft Paper on Online Reputation. As is evident in the case study and will be revisited in the analysis with a feminist bioethical framework on vulnerability, there are too many opportunities for the consent of individuals to be either overlooked or disregarded. 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.

6. APPLYING THE CASE STUDY: An analysis of the BMHS application through the review of a feminist bioethical framework of vulnerable populations, the Right to Explanation, and the Right to be Forgotten.

As has been demonstrated in this paper, the collection of data and its use in opaque algorithms is problematic for society, particularly for those members of society that have more exposure to the risks of the embedded biases within them due to their own vulnerabilities. Accordingly, Frank Pasquale declares that, "runaway data isn't only creepy. It can have real costs. Scoring is spreading rapidly from finance to more intimate fields. Health scores already exist, and a "body score" may someday be even more important than your credit score. Mobile medical apps and social networks offer powerful opportunities to find support, form communities, and address health issues. But they also offer unprecedented surveillance of health data, largely ungoverned by traditional health privacy laws (which focus on doctors, hospitals, and insurers). Furthermore, they open the door to frightening and manipulative uses of that data by ranking intermediaries— data scorers and brokers— and the businesses, employers, and government agencies they inform" (26). The reach the systems and their underlying algorithms have now and will continue to have in the future is quite possibly endless.

Hoffman *et al.* outlined the shortcoming identified prior to the BMHS, namely that police officers lacked appropriate training in dealing with "PSMDs" and whether the BMHS application works to address this shortcoming at all. One of the goals of the BMHS was to establish effective integration between front-line services, but this thesis has raised and will address to what extent and costs this goal is acceptable.

The itemized success by Hoffman et al is that the BMHS, "collect[ed] more detailed information on the characteristics of persons who police officers had interaction with," and that

it, “enabled police officers to capture, in much more detail, the characteristics of PSMDs who were hospitalized but were not based on diagnosis” (Hoffman, 32). This raises the question as to why collecting personal information of private and vulnerable citizens was deemed as a successful results of the pilot project. Furthermore, the clearly articulated limitations of the study — where, as previously mentioned, the sample for the study was, “biased towards white […] population with almost no representation towards First Nations” (33) and the lack of, for example, the ability to measure inter-rater reliability — raises questions surrounding the relevance of a pilot project that will have such a dramatic effect on the personal privacy of a group of individuals who are already more subjected to vulnerabilities to begin with.

Whether Hoffman et al’s pilot study and the resulting digital application would have been even allowable or socially acceptable if the sample population (“PSMDs”) had had the ability to remove their consent or refute the sharing of their personal information is clear. Their vulnerabilities made them further vulnerable as test subjects for the pilot. As Sisti and Stramondo assert, the circular problem of consent made these members of society more available as subjects because of the limitations to consent generated by their vulnerabilities.

In reviewing the BMHS in relation to the previously presented articles on vulnerability, which outline its problematic relationship with consent, it is clear that the BMHS falls short in a number of ways in both adequately addressing PSMD’s as anything more than a problem that needs to be solved, as opposed to individuals who are lacking in their autonomy and ability to work through the factors causing their vulnerabilities. As outlined in Chapter 3 of this thesis, there are three main areas to consider on how the BMHS application affects the autonomy of the persons with vulnerabilities that come into contact with it. As previously outlined, they are: 1) the ability or inability of the “PSMD” — or any individual for that matter — to consent to the

collection and later use of their data and personal information; 2) the issues of privacy around the data collected and who has access to it, either immediately or in the future; 3) the sharing of this data beyond the BMHS application, where it will be stored and any future uses of it beyond the intent through which it was originally collected.

a) The ability to consent to the collection and later use of their data and personal information

The BMHS neither requires the consent to collect the data and information of the individual approached by police officers nor does it address their privacy concerns explicitly in the HealthIM privacy policy. This, as per Sisti and Stramondo, is problematic as it further reinforces their vulnerability, namely through the shame it can further instill. If we are all oppressed in some way, that would imply that we are all incompetent and therefore unable to volunteer ourselves for consent. In this logic, if we are all vulnerable in some way, that would imply that we are all unable to be autonomous. Echoing the feminist bioethical theory on vulnerable populations, variants between individuals should be considered for all policies in order to include each individual. Not doing so would be problematic because it would effectively eliminate every person from participating in our society. Emphasizes Pasquale, "The power to include, exclude, and rank is the power to ensure that certain public impressions become permanent, while others remain fleeting" (14).

By not requiring the consent of the individuals the BMHS collects data and information on, it operates in a procedural-based approach to informed consent. Even if consent were required from the individuals whose data is collected by the BHMS application because of its underlying inability torecognize the individuals as anything more than the sum of their vulnerabilities, it would only be a form of procedural consent, "where patients need only pass the

threshold elements to be considered fully autonomous — considers members of oppressed groups to be procedurally autonomous, while missing the contextual and substantive dimensions of choice that might diminish their autonomy" (Sisti and Stramondo, 72).

One of the risks Rogers *et al.* points to is the risk of paternalistic protections and thereby responding is ways that are too broad and that do too much (12). The BMHS clearly crosses this line by emphasizing vulnerability to harm as the main measurement against which they enable the algorithm to render its decisions. The BMHS works to reinforce the necessity of police intervention, regardless of their lack of training, and reinforces the concept that "PSMDs" require police intervention and not another form of intervention or no intervention at all. This acts to further stigmatization and create paternalism, and, according to Rogers *et al.*, creates the conditions for pathogenic vulnerability (27).

As we are all vulnerable in some way, policies and programs designed to support citizens are an important endeavour, in that, "humans are inevitably dependent on others for our care and nurture at one or more points in our lives" (Rogers *et al.*, 30.) However, as per a feminist bioethical framework, these need to be focused on the central goal of enhancing human agency for all. This cannot be done by focusing on vulnerabilities as defining characteristics of an individual, but rather by seeing the vulnerabilities as pieces of that individual that need to be addressed by social policies, programs or systems that enable their autonomy. "A carefully developed account of vulnerability will assist in understanding the ways in which institutions and practices, including those related to health, shape people's inherent and situational resilience and vulnerabilities; and in developing respectful responses to these vulnerabilities" (Rogers *et al.*, 32).

Ultimately, the issues raised in this case study, specifically those of consent, privacy and autonomy, are indicative of more systemic issues in policies and programs designed to work with and for vulnerable populations. The BMHS is just one of many provincially funded programs working with data collection to address various issues in health care. In this vein, there is a larger problem that needs to be addressed before further public funds and resources are put into programs that increase the exposure and risk to an individual's vital needs. "Namely, as more human characteristics come under that domain of biotechnology's control it will be an increasing challenge to determine whether the enhancements people embrace are voluntarily chosen or are the product of sexist, ableist, racist, classist, and heterosexist standards they have internalized from their oppressive socialization" (Sisti and Stramondo, 79).

b) The issues of privacy around the data collected and who has access to it, either immediately or in the future;

The Right to Explanation, as introduced by Kaminski in Chapter 5 of this thesis, provides an individual with protections surrounding the use of their data. As previously stated, this provides an individual a level of transparency through their data's potential use, present or future, and installs provisions supporting an individual's autonomy throughout the use of their data and information. These provisions, such as the right to a human reviewer, work to ensure their privacy throughout the lifetime of any data that may be collected and used in algorithms and ADMSs.

In the BMHS, the privacy clauses around the application do not provide the individual being encountered by police the ability to understand how their data will be used, where it will be used, or how it may be used in the future. This is dangerous because while an individual is

determined “at-risk” by an officer or the BMHS algorithm, there are two nefarious systems at play that not only remove the need to get the individuals consent but also the need for them to consent to their data and information being used against them, either immediately following the interaction or much further in the future. “We encounter two foundational interpretations that establish the measurable type of ‘at-risk.’ One, there’s an explicit value judgment that social connections themselves are the most efficacious indicators of ‘at-risk’ behavior. And two, there’s an implicit value judgment about what data is available, and eventually used, in processing” (Cheney-Lippold, 483). To this point, there is a possibility that, should there not be any controls over how, when or where the data is used in the future, this individual might never be able to remove the label, and therefore stigma, of being “at-risk”. This then means that they may never be more than a “PSMD”, or a person in a mental health crisis. This further holds that, without a Right to Explanation, they will forever be defined by this moment in time. An obstacle is therefore created to the individual’s ability in exercising choice and enabling their autonomy as they will be subjected to stereotypes based on this moment in time forever, establishing the ADMSs as an invisible tool of discrimination in this way.

In his book *The Black Box Society,* Pasquale outlines the importance of respecting the origin of data through its tracking and the need for auditing systems to enforce and support the necessary tracking. “Tracking data sources should also help individuals correct mistakes […] When the follow-on users of bad data don’t know where it came from, they may not believe the data subject. If they kept track of the provenance of their data, the process of correction would be easier. […] First, as data used intensifies, it will be hard for persons (even with the air of new software and professional help) to keep track of exactly where and how they’re being characterized. Second, in many contexts, even accurate, true data can be unfairly or

discriminatorily deployed." (146.) The RtE and its provisions support the necessary steps in data provenance that Pasquale calls for and the protections helping to enable autonomy that feminist bioethical theorists would support in their identified needs for persons with vulnerability. It enables an individual to be defined by more than a moment in time. It further enables that individual to have access to their data and have it reviewed by a human reviewer. It does not allow for companies to hide behind the argument of trade-secrets and gives an individual the right to see how they were portrayed behind the opaque veil of the algorithm, ensuring the provision of a level of transparency and guarding them, to a degree, against profiling techniques. Added to this, it ensures that an individual has the information they need to be able to contest a decisions, correct inaccuracies or request erasure, as we will see in the next section reviewing the BMHS application and the RtbF.

c) the sharing of this data beyond the BMHS application, where it will be stored and any future uses of it beyond the intent through which it was originally collected.

As introduced earlier in this thesis by Bartolini and Siry, the RtbF is another arm of Data Subject protection which is found in the GDPR. Plainly, the RtbF is the ability for a DS to regain control over their own data after the data has been collected. It specifically addresses the issues of consent and whether or not it can be, in effect, retroactively removed after its initial granting. Not doing so enables our data to speak for itself, with potentially dangerous repercussions. As stated by Cheney-Lippold, "Who we are in the face of algorithmic interpretation is who we are computationally calculated to be. And like being an algorithmic celebrity and/or unreliable, when our embodied individualities get ignored, we increasingly lose control not just over life but over how life itself is defined" (169).

Bartolini and Siry outlined how the DS should maintain a degree of control over their own data, as well as any current or future processing of that data. Involving the DS in a relationship with the data controller would increased and enable their autonomy, empowering them through this relationship. As per a feminist bioethical framework on vulnerability, this is consistent with accepting the DS as more than the sum of their vulnerabilities. By reinforcing the control over their data, that individual's autonomy, in their ability to make an informed decision according to their own reasonings and removed from negative external forces (Christman), is enabled and supported.

Through addressing the issues raised by Bartolini and Siry, an individual's autonomy, regardless of their ability to give consent at the time their data is collected, is supported through the RtbF. Earlier mentioned issues such as storage (dynamic versus static) and the problems surrounding third-parties using data without receiving explicit consent can be better addressed through the RtbF, moreso than they are under current Canadian legislation. Enabling the DS to be able to remove their consent after the fact puts the onus on the users of the data to track and control where the data they have access to comes from. It allows an individual, regardless of the number of vulnerabilities they are exposed to, to be more prudent over when and where their data is used. It also, in the vein of Sisti and Stramondo, would work to support the future abilities of that individual to understand what consent is and what it means to them.

In the BMHS, it is clear through the evidence presented in this thesis that consent is not required from the individual whose data is being collected for use, whether taken from that individual while in a crisis situation or not. It is clear that this data can be stored without respect for the data subject – who in this case is the police officer and not even the individual whose data is being collected at a point in their lives when they are most vulnerable. As outlined earlier, the

privacy policy of the BMHS addresses the destroying of information upon request, however acknowledges that one copy of the information collected would be retained and that there are even further limitations in their ability to do so. The RtbF works to protect all individuals in situations like these with the understanding that protection of an individual's data is more than just bits and pieces of data but is the property of that individual. The protection of rights of the individual and the obligations of the data processor, as discussed in this thesis, are critical to determining the ownership of the data and the regulations that must be made around its use. The ownership of data is clarified through the RtbF and enables the autonomy of all individuals, particularly for those individuals with vulnerabilities who might not have been even able to give consent in the first place.

7. CONCLUSION: Why the RtE and the RtbF are, or need to be, critical pieces of privacy and data ethics legislation as they foster autonomy.

Ultimately, if the goal of the BMHS was to collect data on PSMDs and to share it across front-line agencies, then the application is successful. However, if the goal of the application was to prepare police officers and other front-line workers to better aid and assist individuals with increased vulnerabilities in a way that would protect or enable their autonomy, recognizing their vulnerabilities and asserting their needs through the requirement of their consent, then the application failed. A feminist bioethical framework for vulnerability and consent would call for and require the latter. Pasquale, calling for reform, says:

> the more the black boxes of corporate practices in the areas are revealed, the more pressure will mount to change them. What might real reform look like? When it comes to reputation, it would mean focussing less on trying to control the collection of data up front, and more on its use – how companies and government are actually deploying it to make decisions (140-1).

Data collection does not have to be the proverbial boogey-man, where only bad things can result from it. With protections like the RtE and the RtbF, which as demonstrated can work towards ensuring that autonomy 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. It is the ethical considerations and removal of the blank cheque given to these programs that is problematic.

If we cannot put ethics inside the black boxes – due to, as discussed earlier in this thesis, feedback loops, restricted available data sources and the inherent biases the humans designing the algorithms hold – then we need to build ethics around the black boxes. In particular, in designing algorithms and ADMSs that will be working with vulnerable populations, ethical considerations, particularly those involving persons with increased vulnerabilities, need to be addressed. This would be accomplished through the application of a feminist bioethical framework on vulnerability, which would in turn, I believe, be supportive of the inclusion of the Rights to Explanation and to be Forgotten, as they both enable individual autonomy through consent, both during and after an individual's data is collected, stored and used.

Ultimately, as has been demonstrated through this thesis, there are biases and judgements encoded in ADMSs that perpetuate the vulnerabilities of marginalized individuals. As a result, programs that are designed to assist these individuals, like the BMHS, actually work against them through paternalistic protections and instead work in favour of the existing power structure and those who are subjected to fewer vulnerabilities. These encoded biases further exacerbate the divides and work to reinforce the haves and have-nots within our society. Asserts Pasquale, "what we do know is that those at the top of the heap will succeed further, thanks in large part to the reputation incurred by past success; those at the bottom are likely to endure cascading disadvantages. Despite the promises of freedom and self-determination held out by the lords of the information age, black box methods are just as likely to entrench a digital aristocracy as to empower experts."

In providing more transparency and therefore autonomy in the operations of ADMSs through the appropriate interpretation and requirement of consent, both at and after the point of data collection, the benefits these systems purport to hold could be better realized and supported. "Open uses of technology hold a very different kind of promise. Instead of using surveillance technology against American citizens, the government could deploy it on our behalf, to monitor and contain corporate greed and waste. Public options in technology and finance would make our social world both fairer and more comprehensible. Rather than contort ourselves to fit "an impersonal economy lacking a truly human purpose," we might ask how institutions could be reshaped to meet higher ends than shareholder value" (Pasquale, 217-8).

There is a propensity in our society to always be looking for the new advancement that will enhance our lives, our jobs, 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 work to relegate all of us to pieces of data to be faster and more easily consumed, judged and controlled. Asserts Pasquale:

> Capitalist democracies increasingly use automated processes to assess risk and allocated opportunities. The companies that control these processes are some of the most dynamic, profitable, and important parts of the information economy. All of these services make use of algorithms, usually secret, to bring some order to vast amounts of information. The allure of the technology is clear – the ancient aspiration to predict the future, tempered

> with a modern twist of statistical sobriety. Yet in a climate of secrecy, bad information is as likely to endure as good, and to result in unfair and even disastrous predictions. This is why the wholesale use of black box modeling, however profitable it is for the insiders who manage it, is dangerous to society as a whole. It's bad enough when innocent individuals are hurt, branded as security threats or goldbrickers or credit risks by inaccuracies they can't contest and may not even know about. Modeling is even worse when unfair or inappropriate considerations combine with the power of algorithms to create the failures they claim to merely predict (216-7).

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. This framework would be best designed through a feminist bioethical lens in that these theories focus on consent and its effects on autonomy. Two of the steps towards a feminist bioethical 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.

WORKS CITED

Bartolini, and Siry. “The RtbF in the Light of the Consent of the Data Subject.” Computer Law &Amp; Security Review: The International Journal of Technology Law and Practice, vol. 32, no. 2, 2016, pp. 218–237.

Boddington, Paula. *Towards a Code of Ethics for Artificial Intelligence*. Springer International Publishing.

Canada, Office of the Privacy Commissioner of. Bill S-4, An Act to Amend the Personal Information Protection and Electronic Documents Act and to Make a Consequential Amendment to Another Act (the Digital Privacy Act) - February 12, 2015. 17 Feb. 2015, https://www.priv.gc.ca/en/opc-actions-and-decisions/advice-to-parliament/2015/parl_sub_150212/.

Canada, Office of the Privacy Commissioner of. Consent and Privacy - A Discussion Paper Exploring Potential Enhancements to Consent under the Personal Information Protection and Electronic Documents Act. 11 May 2016, https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2016/consent_201605/.

Canada, Office of the Privacy Commissioner of. 2017-18 Annual Report to Parliament on the Personal Information Protection and Electronic Documents Act and the Privacy Act. 27

Sept. 2018, https://www.priv.gc.ca/en/opc-actions-and-decisions/ar_index/201718/ar_201718/.

Canada, Office of the Privacy Commissioner of. Guidelines for Obtaining Meaningful Consent. 24 May 2018, https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gl_omc_201805/.

Canada, Office of the Privacy Commissioner of. Online Reputation - What Are They Saying about Me? 21 Jan. 2016, https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2016/or_201601/.

Canada, Office of the Privacy Commissioner of. Guideline on Inappropriate Data Practices: Interpretation and Application of Subsection 5(3). 24 May 2018, https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gd_53_201805/.

Canada, Office of the Privacy Commissioner of. *Privacy by Design*. 8 Oct. 2007, https://www.priv.gc.ca/en/blog/20071008/.

Christman, John, "Autonomy in Moral and Political Philosophy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>.

Cheney-Lippold, John. We Are Data: Algorithms and The Making of Our Digital Selves. NYU Press. Kindle Edition.

"Declaration of Montréal for a Responsible Development of AI." *Declaration of Montreal;AI for a Responsible Development of AI,* https://www.montrealdeclaration-responsibleai.com. Accessed 27 Apr. 2019.

"Does Canadian Privacy Law Apply to Google Search?" Michael Geist, 16 Oct. 2018, http://www.michaelgeist.ca/2018/10/does-canadian-privacy-law-apply-to-google-search/.

"Dominated by Men." Reuters, https://fingfx.thomsonreuters.com/gfx/rngs/AMAZON.COM-JOBS-AUTOMATION/010080Q91F6/index.html. Accessed 24 Feb. 2019.

Donchin, Anne and Scully, Jackie, "Feminist Bioethics", *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2015/entries/feminist-bioethics/>.

Frankish, Keith, and William M. Ramsey, editors. The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, 2014.

Google Loses "right to Be Forgotten" Case. 13 Apr. 2018, https://www.bbc.com/news/technology-43752344.

Granville, Kevin. "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens." The New York Times, 14 May 2018, https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

HealthIM (1). https://healthim.com/assess. Accessed 24 Feb. 2019.

HealthIM (2). https://healthim.com/discover. Accessed 24 Feb. 2019.

HealthIM (3). Privacy Policy. https://healthim.com/privacy-policy.pdf. Accessed 24 Feb. 2019.

Hoffman, R., *et al.*, The use of a brief mental health screener to enhance the ability of police officers to identify persons with serious mental disorders. International Journal of Law and Psychiatry (2016), http://dx.doi.org/10.1016/j.ijlp.2016.02.031

Hogenboom, Melissa. *Locked up and Vulnerable: When Prison Makes Things Worse.* http://www.bbc.com/future/story/20180411-locked-up-and-vulnerable-when-prison-makes-things-worse. Accessed 27 Apr. 2019.

Israni, Ellora Thadaney. "When an Algorithm Helps Send You to Prison", NY Times, Oct. 26, 2017. https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

Kaminski, Margot E., The RtE, Explained (June 15, 2018). U of Colorado Law Legal Studies Research Paper No. 18-24. Accessed at: https://ssrn.com/abstract=3196985

Lacey, Hugh. Is Science Value Free? : Values and Scientific Understanding. 1999. Web.

Lamb, Creig. "The Talented Mr. Robot: The impact of automation on Canada's workforce", Brookfield Institute. June 2016. http://brookfieldinstitute.ca/research-analysis/automation/

Larson, Jeff, *et al.* "How We Analyzed the Compas Recidivism Algorithm." ProPublica, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed: Feb 5th, 2018

"MCRRT." St. Joseph's Healthcare Hamilton, https://www.stjoes.ca/health-services/mental-health-addiction-services/mental-health-services/coast/mcrrt?resourceID=5744. Accessed 24 Feb. 2019.

Mobile Rapid Response Team | Hamilton Police Service. https://hamiltonpolice.on.ca/prevention/mental-health/mobile-rapid-response-team. Accessed 24 Feb. 2019.

Ontario, OCHIS: Office of the Chief Health Innovation Scientist, May 7, 2018. http://health.gov.on.ca/en/pro/programs/ochis/value-based-innovation-program.aspx

O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown/Archetype. Kindle Edition.

"Our Principles." *Google AI,* https://ai.google/principles/. Accessed 27 Apr. 2019.

Pasquale, Frank. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press. Kindle Edition.

"Recital 71 - Profiling." General Data Protection Regulation (GDPR), https://gdpr-info.eu/recitals/no-71/. Accessed 24 Feb. 2019.

Rogers, Wendy, *et al.* "Why Bioethics Needs a Concept of Vulnerability." *International Journal of Feminist Approaches to Bioethics*, vol. 5, no. 2, 2012, pp. 11–38. *JSTOR*, JSTOR, www.jstor.org/stable/10.2979/intjfemappbio.5.2.11.

Sisti, Dominic A. "Competence, Voluntariness, and Oppressive Socialization: A Feminist Critique of the Threshold Elements of Informed Consent." *IJFAB: International Journal of Feminist Approaches to Bioethics*, vol. 8, no. 1, 2015, pp. 67–85.

State v. Loomis. https://harvardlawreview.org/2017/03/state-v-loomis/. Accessed 27 Apr. 2019.

The Cambridge Handbook of Artificial Intelligence. Cambridge University Press. Kindle Edition.

"The Google Inc. v. Equustek Solutions Inc. Decision." Gowling WLG, http://gowlingwlg.com/en/insights-resources/articles/2017/google-inc-v-equustek-solutions-inc-decision/. Accessed 25 Feb. 2019.

Van Duin, Stefan, Bakhshi, Naser. "Part 1: Artificial Intelligence Defined: The most used terminology around it", Deloitte. Published March 2017. https://www2.deloitte.com/se/sv/pages/technology/articles/part1-artificial-intelligence-defined.html

Winston, Morton Emanuel., and Edelbach, Ralph. *Society, Ethics, and Technology*. 4th ed., updated.. ed., Wadsworth, Cengage Learning.