



# THE SCHOOL OF PUBLIC POLICY

---

## MASTER OF PUBLIC POLICY CAPSTONE PROJECT

Governing Artificial Intelligence: Opportunities in the Canadian Public and Private Sectors

**Submitted by:**

Thomas Mathieu

**Approved by Supervisor:**

Supervisor name and date supervisor approved and signed form here

Submitted in fulfillment of the requirements of PPOL 623 and completion of the requirements for the Master of Public Policy degree

## **Acknowledgements**

I first thank my parents, Benoit and Tanya, and to my grandparents, Bill and Ursula, for their continuous support and encouragement during the past year. I would not have been able to complete the degree without your help, and all of you cheering me on. Thank you.

Next, thank you to my partner Viktoria Grynenko. I don't know what I would have done without you. When the going was hard, I relied on your support and your caring to continue. Thank you.

Lastly, thank you to my supervisor Jean-Christophe Boucher for your guidance and thoughtful advice in completing this project. I appreciated your focused and practical help as well as your understanding. Thank you.

**Contents**

Introduction ..... 1

Section 1: What is AI? ..... 2

    AI – The Technical Dimension ..... 2

    AI – The Social-Technical Dimension ..... 5

Section 2: Perspectives on AI Governance ..... 11

    Principles and Ethics ..... 11

    Human Rights ..... 14

    Impact Assessments and Accountability ..... 16

Section 3: Approaches to AI Governance ..... 19

    The European Union ..... 19

        EU Proposal for regulations on AI ..... 19

        General Data Protection Regulation (GDPR) ..... 24

    Canada ..... 25

        Directive on Automated Decision Making ..... 25

        Digital Charter Implementation Act, 2022 ..... 28

Section 4: Analysis of AI Governance in Canada ..... 31

    Directive on Automated Decision Making ..... 31

    Digital Charter Implementation Act, 2022 ..... 33

Conclusion ..... 35

    Directive on Automated Decision Making ..... 37

    Bill C-27 Digital Charter Implementation Act, 2022 ..... 38

References ..... 39

## Introduction

In Canada, AI has been used to improve government services. A pilot study was conducted in 2018 on using AI to surveil social media posts to monitor for indicators of potential suicide (Canada 2018). More recently in May 2022, the Department of Citizenship and Immigration investigated employing AI to triage visa applications in a project called *Advanced Analytics Triage of Overseas Temporary Resident Visa Application* (Canada 2022g).

AI is here, but this means that the challenges associated with the use of AI are here too. One common challenge with AI systems are biased outputs. Biased outputs occur when AI systems are developed using data that reflects biased social phenomena or with incomplete data, and then reflects those biases in their outputs (Bolukbasi et al. 2016; Danks and London 2017). In 2020, it was demonstrated that police services across Canada were using predictive algorithms and AI as part of their policing (Citizen Lab 2020, 2). Building off historical crime data, a predictive policing algorithm outputs an estimate of where crime is likely to occur (Hiller 2021,1). Because these algorithms are based on historical policing data in which police target minorities and low-income groups, the algorithms flag areas where minorities and low-income groups live as ‘high crime’ areas (Hiller 2021, 2). This results in over-policing of those vulnerable groups. The further over-policing may impact the human rights of individuals in the areas flagged as ‘high crime’ because police may “...require less suspicious behaviour from the detainee to justify detention than if they had been in area that wasn’t flagged as “high crime” (Hiller 2021, 2).

In Canada, regulatory governance of AI is lacking. There is no legislation that specifically regulates the development or application of AI in the private sector. Bill C-27, currently before the House of Commons, is an important step towards regulating AI in the private sector. AI systems

used in the federal public sector are regulated through a Treasury Board of Canada Secretariat Directive. The Treasury Board Directive requires government departments complete an impact assessment on their AI system before it can come into service. This capstone will show that the private sector and public sector governance of AI in Canada are held back from effective governance of AI by missing key regulatory elements that address the challenges of AI. How can the Government of Canada regulate Artificial Intelligence in the public and private sectors in a way that addresses challenges associated with its use, but does not stifle the benefits of Artificial Intelligence?

This capstone is divided into four main sections. The first section will provide an overview of the technical elements associated with AI, explore how AI is both technical and social in nature, and lay out several challenges associated with the use of AI. The second main section will provide an overview of three normative approaches to AI governance: ‘ethics guidelines’, ‘human rights’, and ‘impact assessments. It will also show how these three normative approaches are interconnected. The third section will lay out current and pending approaches towards AI governance in the European Union and in Canada. The fourth section will analyze the Government of Canada approach to regulating AI. Finally, the capstone will conclude with three recommendations on how Canada can achieve effective AI governance.

## **Section 1: What is AI?**

### **AI – The Technical Dimension**

The term “artificial intelligence” was used for the first time in a 1955 proposal for a study on how computers could be used to “...solve kinds of problems now reserved for humans” (Horvitz 2017). In recent years an increase in the availability of computational power and the increasing

availability of massive datasets has fueled the growth of AI (Smith and Neupane 2018, 27). The rapid growth of AI is mirrored in the rapid growth in investment into AI firms seen over the past 10 years (CBinsights 2016).

Defining “Artificial Intelligence” (AI) is difficult. The term AI is used to refer to both the research field into AI as well as software programs and capabilities (Nisa, Jacobs, and Villeneuve 2018, 4). All existing AI programs are considered ‘narrow AI’ because they can “...facilitate individual, repetitive tasks by learning from patterns found in data” (Nisa et al. 2018, 4). One general definition of AI is as “...an area of computer science devoted to developing systems that can be taught (e.g., through encoding expert knowledge) or systems that can learn (from data) to make decisions and/or predictions within specific context” (Smith and Neupane 2018, 25). Another general definition is that AI can be thought of “...as an advanced form of data analysis in which a software algorithm can draw conclusions based on the results of the analysis, then act on that conclusion, thus behaving in an “intelligent” fashion” (Smith and Neupane 2018, 27).

Several techniques are commonly used to develop AI systems. ‘Machine learning’ enables AI systems to learn “...by taking in data from existing datasets or through feedback from interaction with their environment” (Smith and Neupane 2018, 31). Machine learning is accomplished through a machine learning algorithm which, over time, can improve its performance in data analysis and analytical modelling (Nisa et al. 2018, 4). There are three general types of machine learning algorithms, but all machine learning models are trained using data (Gebru et al. 2021, 1).

*Supervised learning* algorithms “...are successively fed example data; for each example, the algorithm provides a response (an output or prediction)” (Smith and Neupane 2018, 31). Then, the “...algorithm learns by adjusting its internal parameters such that its response for that particular

example will be more accurate the next time” (Smith and Neupane 2018, 31). The dataset used to train the algorithm is called the ‘training set’. For example, the training set for an image recognition AI system may contain digital images of fruits and each image will be tagged with metadata label indicating the correct answer (Smith and Neupane 2018, 103). Video suggestions on streaming platforms and facial recognition are examples of AI systems developed through supervised learning (Lemelin-Bellerose 2019, 2). *Unsupervised learning* algorithm is “...fed datasets without any associated outputs; the algorithm extracts patterns...from the data” (Smith and Neupane 2018, 4). Unsupervised learning is used to develop virtual assistants such as Siri or Alexa (Lemelin-Bellerose 2019, 2). If an unsupervised algorithm was fed a dataset of fruit, it may classify the data into two clusters, large fruit, and small fruit (Smith and Neupane 2018, 104). The last type is *reinforcement learning* which trains algorithms with either positive or negative feedback on specific actions (Smith and Neupane 2018, 4). Reinforcement learning is often used for developing AI systems with goal-oriented tasks like learning a game or manipulating an object (Smith and Neupane 2018, 4).

Like other contemporary technologies, AI creates policy problems and governance challenges for governments. These challenges are due in part to governments being surprised by the speed of technological progress and lacking tools to intervene and prevent harm to citizens (Whittlestone and Clark 2021, 3). Whittlestone provides several examples of policy challenges related to AI that governments have encountered:

- The rapid deployment and broad diffusion of facial recognition capabilities into the world via chiefly private sector actors (e.g., Clearview AI).
- The development of technologies for better editing and manipulation of videos via a cluster of AI techniques colloquially known as ‘deepfakes’.

- Harmful biases displayed by deployed computer vision and natural language processing AI systems.
- Radicalization of populations through increasingly effective and opaque recommendation systems deployed on platforms such as YouTube, Facebook, and so on (Whittlestone and Clark 2021, 4).

This list masks what several authors term the ‘specification dilemma’ with AI systems. Moss defines the ‘specification dilemma’ as the ability for an AI or algorithmic system to “...cause harm when they fail to work as specified – i.e., in error – but may just as well cause real harms when working exactly as specified” (2021, 5). For example, AI systems for facial recognition cause harm and create policy problems both when they are working in error and when they are working as intended. Facial recognition technologies suffer from differing levels of accuracy, an unintended error, for different demographic groups and this produces harms such as wrongful arrest (Moss et al. 2021, 5). Facial recognition also produces harms such as the limiting freedom of assembly and free association when it is working exactly as intended (Moss et al. 2021, 6).

## AI – The Social-Technical Dimension

The characterization of AI systems and algorithms given in the previous section understands them as “...basically instructions fed to a computer to solve certain problems” (Goffey 2008, 16). This is a technical characterization and places algorithms as the “core stuff” of computer science (Seaver 2017, 1). However, anthropologists and critical scholars, such as Seaver and Dourish, have argued that algorithms are not simply technical objects. What algorithms are exactly and how they relate to the culture in which they are developed is complex. Dourish argues that “...the limits of the term algorithm are determined by social engagements rather than by

technological or material constraint” (Dourish 2016, 3). What this means, according to Seaver, is that “...different people, in different historical moments and social situations have defined algorithms, and their salient qualities differently” (Seaver 2017, 2). He illustrates this claim with the following example:

A data scientist working at Facebook in 2017, a university mathematician working on a proof in 1940, and a doctor establishing treatment procedures in 1995 may all claim, correctly, to be working on “algorithms,” but this does not mean they are talking about the same thing (Seaver 2017, 2).

Seaver distinguishes two positions that can be taken to resolve the complexity. One is ‘algorithms in culture’. This view “...hinges on the idea that algorithms are discrete objects that may be located within cultural context or brought into conversation with cultural concerns” (Seaver 2017, 4). Under this view, an algorithm can alter flows of cultural material such as what films are recommended by a recommender algorithm, and the algorithm can be shaped by culture such as when cultural biases become embodied by the algorithm (Seaver 2017, 4). The other view is ‘algorithms as culture’. Under this view algorithms are not purely technical objects, they are socio-technic objects that “...are unstable objects, culturally enacted by the practices people use to engage with them” (Seaver 2017, 5).

Seaver cites a 2014 study to illustrate the view of ‘algorithms as culture’. In the study, Devendorf and Goodman examined algorithmic interactions on a dating site (Devendorf and Goodman 2014). Seaver summarizes their findings as follows:

Devendorf and Goodman found various actors enacted the site’s algorithm differently: engineers tweaked their code to mediate between the distinctive behaviors of male and female users; some users tried to game the algorithm as they understood it, to generate more desirable matches; other users took the algorithm’s matches as oracular pronouncements, regardless of how they had been produced. No inner truth of the algorithm determined these interactions, and non-technical

outsiders changed the algorithm's function: machine learning systems changed in response to user activity, and engineers accommodated user proclivities in their code. (Seaver 2017, 4)

For Seaver, the study of algorithms on a dating site illustrates that algorithms are produced and interacted with in a “loosely coordinated confusion” and that the public and critics ought to keep that in mind (Seaver 2017, 3).

The socio-technic view of algorithms and AI as culture highlights several challenges with AI. The first challenge is the different types of bias that can be exhibited by AI systems. AI systems, like all computer systems, “...are designed and developed by humans and trained on data generated by humans...” so “...these systems incorporate the biases of their designers and programmers and of the larger culture in which they are created” (Smith and Neupane 2018, 59). It is possible for bias to enter AI systems during the design phase because the implicit values of the people developing a systems influence “...both the selection of applications for development and the features of those applications” (Smith and Neupane 2018, 60). Smith and Neupane argue that this design bias is reflected in the gender distribution of AI application such as chatbots and digital assistants (2018, 60). Data suggests that while chatbots are spread approximately equally between female, male, and genderless characterizations; digital assistants are 67% female and 33% genderless (Coren 2017).

‘Learned bias’ or ‘algorithmic bias’ occurs when AI systems are developed using data that represents biased social phenomena or incomplete data, and then reflect those biases in their outputs (Bolukbasi et al. 2016; Danks and London 2017). An AI system that is trained using an incomplete dataset is likely to perform poorly when given an example that falls outside the scope of its training (Smith and Neupane 2018, 63). This type of bias has been documented in algorithms designed for facial analysis. For example, in a beauty contest judged by an AI system, the 44

winners selected by the algorithm were nearly entirely white because the dataset used to develop the algorithms did not feature minorities or people of colour (The Guardian 2016). So, even though the developers did not deliberately build an algorithm that treated white skin as a sign of beauty, by using a dataset with virtually only white faces the developers effectively created an algorithm that did just that (The Guardian 2016). Another example has to do with clinical trials. Because participants in clinical trials tend to be white males, "...the findings of these studies do not generalize well across the broader population, with outcomes potentially compromised for individuals not represented in the research" (Smith and Neupane 2018, 63). So, "...AI algorithms trained on this data are biased and potentially dangerous to those population that have been excluded" (Smith and Neupane 2018, 63-64).

Algorithms developed from biased social data also have the potential to be biased (Geburu et al. 2021, 1). For example, 'word embeddings' are a common technique utilized in a variety of algorithmic applications in which algorithms and AI systems learn to represent words with similar meaning similarly across diverse texts (Machine Learning Master 2019). Word embeddings form the basis of algorithms with applications for web searching to analyzing curriculum vitae (Bolukbasi et al. 2016, 1). Bolukbasi and colleagues identified that embeddings also pinpoint implicit sexism and gender bias in text (2016, 1). The study's authors developed word embeddings based on Google News articles and showed that the word embeddings exhibited "...female/male gender stereotypes to a disturbing extent" (Bolukbasi et al. 2016, 1). For example, the system that the authors developed would "...offensively answer "man is to computer programmer as women is to x" with x=homemaker" (Bolukbasi et al. 2016, 3). Any system that uses the word embedding is likely to propagate and amplify this bias (Buolamwini and Geburu 2018, 1).

The justice system and policing feature two common examples of biased algorithmic systems that result from socially biased data. First, it has been demonstrated that risk assessment algorithms in the United States are biased against black defendants compared to white defendants (ProPublica.org 2016). These algorithms provide a score representing the risk of re-offense based on the answers to a variety of questions. The investigation found that:

- “The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.”

(ProPublica.org 2016)

It is not difficult to imagine how an algorithm trained on biased historical sentencing and crime data produces outcomes that are biased.

Responses to the ProPublica investigation focused on the particular definition of ‘fairness’ used by the algorithm and whether the algorithm can be described as fair under that definition. However, fairness and the idea of harmful impact on particular groups are social and ethical concepts (Chouldechova 2016, 2). So, a “...risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used” (Chouldechova 2016, 2).

Lastly, even identifying whether there is bias in AI systems is complicated by the nature of some AI systems because the choices and actions of these programs “...can be difficult or impossible to explain” (Smith and Neupane 2018, 66). Certain deep learning techniques, and machine learning techniques used to develop AI create AI with highly complex arrays of millions of encoded patterns (Smith and Neupane 2018, 66). If an AI makes a decision that is unfair to a

person or results in harms to disadvantaged groups, how can that decision be evaluated and how can liability be determined if it is not possible to understand how the decision was reached? The inability to understand some AI decision is often referred to as ‘the black box nature of AI’. The problem of the ‘black box’ nature of AI is compounded by the reluctance of technology firms to allow regulators access to their AI systems which they view as proprietary software (Ananny and Crawford 2018, 985).

## Section 2: Perspectives on AI Governance

The challenges associated with the technical and socio-technic nature of AI has brought about a variety of normative perspectives on how AI should be developed and governed. This section of the capstone will provide an overview of three of these perspectives: principles and ethics, human rights, and impact assessments and accountability.

### Principles and Ethics

The increase in the number of ethical guidelines mirrors the rapid growth seen in the AI sector in recent years (Jobin, Lenca, and Vayena 2019, 3). Private companies, research institutions, public sector organizations, and industry groups have responded to societal fears over problems with AI by publishing documents on what ethical, technical, and other best practices are needed to realize ‘ethical AI’ (Jobin et al. 2019, 3). Despite the abundance of publications by “...seemingly every organization with a connection to technology...” (Fjeld et al. 2020, 3), there is debate over what ethical principles are needed for ‘ethical AI’.

Two studies have attempted to identify consensus in the realms of principle documents. In the first study, Fjeld and co-authors assembled a dataset of 36 guideline style documents that set out principles for “...ethical, rights-respecting, and socially beneficial AI” (Fjeld et al. 2020, 4). The documents were prepared in different cultural environments and are authored by a range of authors including governmental organizations, professional associations, industry, and advocacy groups (Fjeld et al. 2020, 4). The authors identify consensus in the documents around eight key themes.

- Privacy: AI systems should respect the privacy of individuals both in use and development
- Accountability: Accountability for AI systems should be appropriately distributed.

- Safety and Security: AI systems should be safe and perform as intended.
- Transparency and Explainability: AI systems should produce intelligible outputs and information should be provided about when they are being used.
- Fairness and Non-discrimination: AI systems should be designed to promote fairness and inclusivity.
- Human Control of Technology: Important decision should be subject to review by humans.
- Professional Responsibility: Developers of AI systems should uphold professionalism.
- Promotion of Human Values: AI systems should promote humanities well-being and correspond with core values (Fjeld et al. 2020, 4-5).

Of the eight themes, ‘fairness and non-discrimination’ was present in 100% of the documents in their dataset while ‘promotion of human values’ was present in 69% of the documents. An additional finding of the study is that the eight themes may constitute a normative core to AI governance because the more recent the publications tended to cover all eight themes while earlier publications did not (Fjeld et al. 2020, 5).

In the second study, a review of publications on AI principles and guidelines identified 84 such documents (Jobin et al. 2019, 3). The review also identified that the number of publications has increased over time with 88% of publications released since 2016 (Jobin et al. 2019, 3). Breaking down the results showed private companies produced the most documents at 22.6% followed closely by governmental agencies at 21.4% (Jobin et al. 2019, 3). Academic research institutions produced 10.7% of the documents (Jobin et al. 2019, 3). The United States and the European Union are the main regions that produce such documents (Jobin et al. 2019, 6).

After identifying several overarching ethical values and principles, Jobin concludes that no “...single ethical principle appeared to be common to the entire corpus of documents although

there is an emerging convergence around the following principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al. 2019, 7). The authors go on to note that despite the consensus around top level principles, a closer examination shows that there are “...significant semantic and conceptual divergences in...how...the principles are interpreted...” (Jobin et al. 2019, 7). For example, ‘transparency’ which was the most prevalent principle identified in the literature has “...significant variation in relation to the interpretation, justification, domain of application, and mode of achievement” (Jobin et al. 2019, 7-8). According to Jobin, many sources suggest increased disclosure of information by AI developers, but *what* exactly should be communicated varies greatly depending on how the AI system is used, what data is used, the source code of the AI, and limitations of the AI (Jobin et al. 2019, 8).

According to Floridi and Cowls, the array of principles described by the different documents creates the problem of ‘principle proliferation’ (2019, 2). This is the idea that “...the sheer volume of proposed principles threatens to overwhelm and confuse...” (Floridi and Cowls 2019, 2). Beyond that there are two further problems from the super abundance of core principles: “Either the various sets of ethical principles for AI are similar, leading to unnecessary repetitions and redundancy, or, if they differ significantly, confusion and ambiguating will result instead” (Floridi and Cowls 2019, 2). After analyzing several high profile and recent publications on AI principles, the authors identify convergence in the principles of beneficence, non-maleficence, autonomy, justice, and explicability (Floridi and Cowls 2019, 6).

Latonero agrees that the abundance of principles amounts to principle proliferation as defined by Floridi and Cowls, but he argues that principle proliferation “...signals a crisis of legitimacy that could have negative ramifications on a global scale (2020, 2). Because AI does not belong to only one discipline or profession but is instead a cross disciplinary cluster of technologies

no single profession or institution is the authority (Latouner 2020, 2). Therefore, the “...sheer number of principles makes us unsure who to trust, who sets the rules of the road, and who would enforce them” (Latouner 2020, 2). Latouner also shows that there is an abundance of national strategies on AI on top of the guidelines by private corporations with over 30 different nationals from different groups like the EU and the G7 producing national strategies (Latouner 2020, 3).

## Human Rights

One response to principle proliferation and the crisis of legitimacy from the abundance of ethical principles is to use international human rights as a guide for AI development towards a common good (Latouner 2020, 5). For Latouner, human rights provide a “...a language for communicating about social risks and harms grounded in rights like privacy, freedom of expression, assembly, non-discrimination, equality, and dignity” (Latouner 2020, 5). Essentially, human rights can provide a grounding for the plurality of principles needed to guide AI development (Latouner 2020, 7).

Other scholars support human rights as an effective way to influence policy that guides AI development. The ethics statements and AI development guidelines written by private firms and groups are useful for clarifying their positions, but they lack normative sway over other groups, and they do not establish a governance framework all can operate under (Donahoe and Metzger 2019, 116). Donahoe argues that human rights can provide a comprehensive global governance framework for AI because human rights, as a system, already has global buy in and articulates the roles of government and the private sector (Donahoe and Metzger 2019, 118). The *Universal Declaration of Human Rights* and the *International Bill of Human Rights* are the documents that have together created “...a body of international human-rights law that establishes the obligations

of governments to protect (and not violate) the rights of citizens and other people within those governments' respective territories and jurisdictions" (Donahoe and Metzger 2019, 119).

Donahoe describes four features of the human rights framework that she argues make it effective at providing a global framework for AI governance.

1) It puts the human person at the center of any assessment of AI and makes impact on humans the focal point of governance; 2) it covers the wide range of pressing concerns that AI raises, both procedural and substantive; 3) it outlines the roles and duties of both governments and the private sector in protecting and respecting human rights; and 4) it has the geopolitical advantage of resting on a broad global consensus, shared by many countries around the world and understood to be universally applicable. Together, these features make a strong case to support use of this existing framework to guide the governance of AI (Donahoe and Metzger 2019, 119).

Mantelero and co-authors also argue that human rights are an effective framework for guiding the development of AI and regulating its impacts (2021, 1). He criticizes the array of ethical principle publications as lacking clear indications for preferring one ethical framework over another (Mantelero and Esposito 2021, 1). He also argues that ethical statements are necessarily context dependent and that the local nature of ethical values "...undermines the ambition to provide global standards or, where certain values are claimed to have general relevance, may betray a risk of ethical colonialism" (Mantelero and Esposito 2021, 1).

Mantelero criticizes the Fjeld and Jobin studies that attempted to identify consensus in ethical principles across different guideline publications (Mantelero and Esposito 2021, 5). Specifically, he argues that these studies lack "...policy perspective and contextual analysis..." for several reasons (Mantelero and Esposito 2021, 5). First, widely different sources are considered on the same level, regardless of whether they are adopted by a government body, an independent author, or a large firm and this suggests that the mere frequency of values is most important (Mantelero and Esposito 2021, 5). Ultimately, for Mantelero, human rights are a specific

crystallization of and a concretization of ethical values “...that are circumscribed and contextualized by legal provisions and judicial decisions” (Mantelero and Esposito 2021, 4). Human rights then need to be operationalized into tools such as human rights impact assessments for AI to mitigate the potential harms from AI (Mantelero and Esposito 2021, 34).

## Impact Assessments and Accountability

Human Rights Impact Assessments and Algorithmic Impact Assessments are specific applications of the larger governance practice of impact assessments. Impact assessments are a widely undertaken practice and are “...simply defined...the process of identifying the future consequences of a current or proposed action” (International Association for Impact Assessment 2022). In the context of AI, *Algorithmic Impact Assessments* are a governance practice for “...delineating accountability, rendering visible the harms caused by algorithmic systems, and ensuring steps are taken to ameliorate those harms” (Metcalf et al. 2021, 1). All existing impact assessments can be described as having 10 components (Moss et al. 2021, 1). These components are sources of legitimacy, actors and forum, catalyzing event, time frame, public access, public consultation, method, assessors, impact, and harms and redress (Metcalf et al. 2021, 1).

The impacts at the center of Algorithmic Impact Assessments “...are constructs that act as proxies for the often conceptually distinct sociomaterial harms algorithmic may produce” (Metcalf et.al 2021, 2). Impacts are distinct from the actual harms experienced by people. For example, a human rights impact assessment might quantify the impacts on human rights as something abstract such as ‘securing life chances’ while the harm is a stifling of rights (Metcalf et al. 2021, 5). Another example for the difference between impacts and harms can be seen for a hiring algorithm in which

the "...impact might be disparate error rates for men and women within a hiring algorithm; the harm would be unfair exclusion from the job (Moss et al. 2021, 7)"

What is crucial for impact assessments as a governance tool is delineating harms and impacts, and accountability is essential for this delineation. According to Moss, impacts can only become knowable from harms "...within an accountability regime which makes it possible to assign responsibility for the effects of a decision, action, or system" (Moss et al. 2021, 7). This is because the accountability relationships are necessary to "...delimit responsibility and causality, there are no "impacts" to measure; without impacts as a common object to act upon, there are no accountability relationships" (Moss et al. 2021, 8). The crucial relationship between accountability, harms, and impacts is that they are co-constructed, impacts do not come before the identification of the parties responsible for the impacts (Moss et al. 2021, 9). However, what is established as 'accountability' depends on the power structures within a society (Metcalf et al. 2021, 5).

Bovens provides a conception of accountability that is widely used in impact-assessments. He defines accountability as "...a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences" (2007, 450). An 'actor' can be an individual, a group or an organization; and, the 'accountability forum' "...can be an agency, such as a parliament, a court, or the audit office (Bovens 2007, 450). For this conception of accountability, it is crucial that there is the possibility of consequences, otherwise it would be indistinguishable from a non-committal provision of information (Bovens 2007, 451). Transparency is among the prerequisites for accountability but does not itself qualify as accountability because it "...does not necessarily involve scrutiny by a specific forum" (Bovens

2007, 453). An implication of this conception of accountability is that actors when anticipate consequences from negative evaluations by accountability forums, they will adjust their policies to avoid consequences (Bovens 2007, 453).

Accountability requires precise delineation in the context of algorithms and AI systems. First, there are questions on who the actors are for algorithms; ‘who is responsible when an algorithm is working incorrectly?’ is a difficult question to answer because the organization using the algorithm is not necessarily the developer of the algorithm (Wieringa 2020, 2). Because of the socio-technic nature of AI, and its complex development, there are questions about who is accountable among the different roles and levels of actors involved in AI (Wieringa 2020, 3). For example, is it decision makers using the AI, developers creating the AI, or users of the AI (Wieringa 2020, 3)? Ultimately, Wieringa specifies algorithmic accountability as follows:

Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system’s lifecycle. In this accountability relationship, multiple actors (e.g., decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct. As different kinds of actors are in play during the life of the system, they may be held to account by various types of fora (e.g., internal/external to the organization, formal/informal), either for particular aspects of the system (i.e. a modular account) or for the entirety of the system (i.e. an integral account). Such fora must be able to pose questions and pass judgement, after which one or several actors may face consequences. The relationship(s) between forum/fora and actor(s) departs from a particular perspective on accountability (2020, 10).

## Section 3: Approaches to AI Governance

This section of the capstone will present two case studies on AI governance. The first will be an examination of the European Union’s approach to AI governance; the second will be an examination of the Canadian approach to AI governance.

### The European Union

The European Union (EU) has two main policies that govern AI, and the regulatory approach incorporates all three of the governance approaches discussed above. The first regulatory instrument is focused on artificial intelligence and is titled the 2021 European Commission *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts* (hereafter “EU Proposal”). The EU *Proposal* has not yet come into effect as of 2022. The second policy is focused on data security, but it has implications for AI governance. This policy is titled the *General Data Protection Regulation* (GDPR). The EU *Proposal* will be discussed first, followed by the GDPR.

#### EU Proposal for regulations on AI

The European Union began work on regulating AI in the mid 2010s. A 2018 communication titled *Artificial Intelligence for Europe* from the Commission to the European Parliament set out the initial EU strategy on AI. The Communication focused on the societal and economic benefits of AI with the expectation that AI is going to transform the world and society like the steam engine transformed the world (European Commission 2018, 1). The European

strategy on AI as outlined in the document aimed to increase the EU's AI uptake in technology and industry, prepare society for the changes that AI will bring about, and ensure that AI operates within an ethical and legal framework connected to the Charter of Fundamental Rights of the EU (European Commission 2018, 3).

The Council of the European Union responded to the Communication in 2019. The response, in the form of a draft conclusion, was supportive of the strategy outlined in the Communication (Council of the European Union 2019, 3). The Conclusion suggested that AI systems need large amounts of data to be developed but the idea was phrased as a "...need for more secure and high quality publicly and privately held data being available..." (Council of the European Union 2019, 6).

Following the Council Conclusions, a White Paper titled *A European approach to excellence and trust* was launched in 2020. Here, "...the Commission supports a regulatory and investment-oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology" (European Commission 2020, 1). The White Paper notes many of the same societal benefits elsewhere, but more clearly articulates the problems associated with AI namely that it poses risks "...for fundamental rights, including personal data and privacy protection and non-discrimination (European Commission 2020, 10).

In 2020, the European Parliament adopted a recommendation to the Commission on a framework "...of the ethical aspects of artificial intelligence, robotics, and related technologies" (European Parliament 2020, 3). The European Commission definition of 'AI system' in the EU *Proposal* is based on two prescriptive elements. The first prescriptive element to the Commission's

definition is that it should be based on “key functional characteristics of the software, in particular the ability, for a given set of human-defined objectives, to generate outputs such as content, predictions, recommendations, or decisions which influence the environment with which the system interacts, be it in a physical or digital dimension” (European Commission 2021,18). The second prescriptive element to the definition is that it should be “...complemented by a list of specific techniques and approaches used for its development...” and that the list of techniques should be kept up to date (European Commission 2021, 18).

The recommendation carries forward the same commitments to human rights that the other EU documents discuss, and that the *Proposal* is based on. However, in a section titled ‘Principles of the requested regulation’ the EU resolution emphasizes ideas that are not carried forward in the Proposal. These new ideas are a “...right to redress; social responsibility and gender equality in artificial intelligence, robotics and related technologies; environmentally sustainable artificial intelligence, robotics and related technologies” (European Parliament 2020, 3).

The EU *Proposal* regulation has four objectives:

- “Ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;
- Ensure legal certainty to facilitate investment and innovation in AI;
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
- Facilitate the development of a single market for lawful, safe, and trustworthy AI applications and prevent market fragmentation” (European Commission 2021, 3).

The European Commission notes that member states cannot achieve these objectives alone due to the “...nature of AI which often relies on large and varied datasets, and which may be embedded in any product or service circulating freely within an internal market...” (European Commission 2021, 6). Fragmented national rules will not protect people or ensure fundamental rights (European Commission 2021, 6) because AI systems “...can be easily deployed in multiple sectors of the economy and society, including cross border, and circulate throughout the Union” (European Commission 2021, 17).

The EU *Proposal* distinguishes between a low-risk and a high-risk classification of AI, but the EU *Proposal* does not provide a definition of a low-risk AI system. One of the primary factors the Commission uses in classifying an AI system as high-risk is the extent of the AI system’s impact on rights set out in *the Charter of Fundamental Rights of the European Union* (European Commission 2021, 28). Among other rights, the EU *Proposal* mentions “...the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information...right to an effective remedy and to a fair trial, right of defence and the presumption of innocence, right to good administration” (European Commission 2021, 24). The EU *Proposal* also states that the impact of AI on the “...fundamental right to a high level of environmental protection enshrined in the Charter and implemented in Union policies should be considered when assessing the severity of the harm that an AI system can cause...” (European Commission 2021, 24).

Stand-alone high-risk systems pose a high risk to health and safety of persons and have the potential of severe harm to fundamental rights (European Commission 2021, 26). The EU *Proposal* identifies a range of AI systems and uses as high risk. For example, AI systems used to evaluate credit scores, or creditworthiness of persons are high-risk because they impact a person’s

access to financial resources towards essential services such as housing, and AI systems used in migration and border control and migration to detect the emotional state of a person should be classified as high risk (European Commission 2021, 27-28). Additionally, most of the AI uses by law enforcement are defined as high-risk due to the power imbalance and resulting deprivation of fundamental rights. The EU *Proposal* specifically mentions that an AI system not trained with high quality data, and that “...does not meet adequate requirements in terms of its accuracy or robustness or is not properly designed and tested before...being put in service...” may single out individuals in a discriminatory manner” (European Commission 2021,28).

Putting a high-risk AI system into service under the EU *Proposal* is accompanied by compliance with several requirements. The first requirement is that the AI system use high quality data in training, validation, and testing sets (European Commission 2021, 29). These training and tests data sets “...should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system” (European Commission 2021, 29). The high-quality data is meant to protect individuals from discrimination that can result from bias in the AI system (European Commission 2021, 29). The second requirement is that high-risk AI systems should have a degree of transparency to facilitate understanding of the system and should be comprehensible to natural persons (European Commission 2021, 30). The system should not be so complex as to be opaque, and a user of the AI system “...should be able to interpret the system output and use it appropriately.” (European Commission 2021, 30) An extension of this transparency requirement is that the AI systems should have clear and available records of technical documents describing the “...general characteristics, capabilities and limitations of the system, algorithms, data, training, testing, and validation processes used as well as documentation of the relevant risk management system” (European Commission 2021, 30). The third requirement

is that high-risk AI systems be state of the art and “...perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity...” (European Commission 2021, 30). Associated with the accuracy requirement is that the “...level of accuracy and accuracy metrics should be communicated to users” (European Commission 2021, 30).

### General Data Protection Regulation (GDPR)

In response to the arrival of the internet in 1995, the EU passed the *Data Protection Directive* (Casey, Farhangi, and Vogl 2019, 145). In 2018, the EU replaced the *Data Protection Directive* with the *General Data Protection Regulation*. Through a reading that combines Article 13, 14, 15 and 22, the GDPR sets out a ‘right to an explanation’ (Kaminsky 2019, 196).

Article 22(1) states: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (European Union 2016, 46). Casey and co-authors note that the GDPR protections set out in article 22 are supported by protections in articles 13-15 that pertain to the rights of individuals "...whose personal information is directly or indirectly implicated by automated processing techniques" (2019, 156). Broadly, articles 13-15 outline the information that companies need to provide individual when their data is subject to automated processing. Casey and co-authors state that Articles, 13(2)(f), 14(2)(g), and 15(1)(h) together mandate a 'right to an explanation' (2019, 156). The articles state that companies must provide individuals with information on "...the existence of automated decision-making, including profiling referred to in Article 22(1) ...and...meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (European Union 2016, 41). So, at "...a minimum, the protection appears to envisage

a limited right for data subjects to understand and verify the basic functionality of certain automated decision-making systems" (Casey et.al. 2019, 158).

The EU's regulatory approach uses all three of the governance approaches identified above. If the EU *Proposal* were implemented in its current form, it would be an effective form of governance for AI systems. Many of the problems that result from both the technical and socio-technic nature of AI are explicitly addressed by the EU *Proposal* and the GDPR. However, the EU's approach is lacking in defining accountability as it relates to assessing the impacts. Specifically, the proposal does not clearly set out who the actors and forum are for assessing accountability, nor does it lay out in detail any consequences. As of June 2022, the EU appears to be moving ahead with implementing the *Proposal* on AI. Most recently, the EU launched a "regulatory sandbox" in Spain to examine how to operationalize requirements of the proposal and to examine conformity assessments (European Commission 2022).

## Canada

Canada's main policy governing in AI is the 2020 *Canadian Directive on Automated Decision Making* (hereafter "the *Directive*") issued by the Treasury Board of Canada Secretariat. This policy will be discussed first. The second policy on AI is contained in Bill C-27, the *Digital Charter Implementation Act, 2022*.

### Directive on Automated Decision Making

The objective of the *Directive* "...is to ensure that Automated Decision Systems" are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law"

(Canada 2019a, 1). The directive applies only to systems that provide external services (Canada 2019a,1). ‘External services’ are defined in the *Policy on Service and Digital* as a service where the intended client is external to the Government of Canada (Canada 2019b, 1). Before producing and implementing an external Automated Decision System, programs are required to complete an Algorithmic Impact Assessment (Canada 2019a, 1). One of the desired outcomes associated with the Algorithmic Impact Assessments is quality assurance. This goal is meant to minimize unintended data biases that unfairly impact outcomes, and to develop processes for continuous monitoring (Canada 2019a, 3).

The Automated Decision Systems are classified into one of four levels based on the impacts that the decisions will have on individuals (Canada 2019a, 4). All four levels of systems are evaluated on the expected impacts in four capacities: the rights of individuals or communities; the health or well-being of individuals or communities; the economic interests of individuals, entities, or communities; and the ongoing sustainability of an ecosystem (Canada 2019a, 4). An impact level I system is expected to lead to impacts that are “reversible and brief” and that “...will likely have little to no impact” on any of the four categories (Canada 2019a, 4). An impact level II system is expected to have moderate impacts on the categories, and that the impacts are likely to be “reversible and short-term” (Canada 2019a, 4). An impact level III system is expected to have high impacts on the four categories, and the impacts are expected to be difficult to reverse and ongoing (Canada 2019a, 4). An impact level IV decision is expected to have a very high impact on the four categories, and the impacts are expected to be irreversible, and perpetual (Canada 2019a, 4).

The different impact levels are associated with different regulatory requirements. For example, an impact level I Automated Decision Making System does not require peer review, approval for the system to operate, or require that a human being be involved in the decision

(Canada 2019a, 4). A level IV Automated Decision-Making System is associated with extensive requirements. A level IV system requires that a peer review be undertaken by at least two of the following types of experts: qualified experts from within the Government of Canada, qualified experts from faculty of a post-secondary institution, a contracted specialized vendor, an advisor specified by the Government of Canada, or the specifications for the decision-making system must be published in a peer reviewed journal (Canada 2019a, 4-5). Impact level IV systems also require that a plain language notice be issued through all service delivery channels including mail (Canada 2019a, 5). Lastly, impact level IV requirements state that decisions "...cannot be made without having a specific human intervention point during the decision-making process; and the final decision must be made by a human " (Canada 2019a, 5).

The algorithmic impact assessment tool used by Canada is structured as an electronic questionnaire (Canada 2022b). The questionnaire is approximately 80 questions long and designed to be completed in at least 35 minutes (Canada 2022b). Each question is scored in one of five categories, and an increasing score moves respondents up through the four impact levels (Karlin and Corriveau, 2018). The five scoring categories for each question are: Socioeconomic impact, Government impact, Data management practices, Procedural fairness considerations, and Complexity of systems (Karlin and Corriveau, 2018). Some of the questions that the survey asks are: "What is your system doing? It is prioritizing files? Making decisions end-to-end?; What types of decisions are resulting from this system? Is this a critical social program? A permit or licensing program? Food or product safety? Crime or fraud detection? Have you consulted everyone you need to (e.g. legal, HR, IT...)" (Canada 2022a). Furthermore, departments are required to upload the results of a completed algorithmic impact assessments to the Open Government Portal (Canada

2022f). As of July 2022, five impacts assessments have been completed, and all five impact assessments have a score that assigns a classification of impact level II (Canada 2022d).

### Digital Charter Implementation Act, 2022

Legislation on AI governance has been brought before the House of Commons twice. The second instance, and the subject of this section, is Bill C-27, the *Digital Charter Implementation Act, 2022*. Bill C-27 appeared for first reading in June 2022. It is a reworked and expanded version of 2020's Bill C-11, which was also called the *Digital Charter Implementation Act*. Bill C-11 attempted amendments to consumer and personal information protections in order "...to protect the personal information of individuals while recognizing the need of organizations to collect, use or disclose personal information in the course of commercial activities" (Canada 2022e). Bill C-11 died on the order paper when the 2021 federal election was called.

The 2020 DCIA and the 2022 DCIA contain identical provisions on algorithmic transparency and a right to an explanation concerning decisions made by an automated decision system. In both versions of the DCIA, the right to an explanation can be found in sections 62 and 63. Section 62 states that an organization must provide "...a general account of the organization's use of any automated decision system to make predictions, recommendations or decision about individuals that could have a significant impact on them" (Bill C-11 2020, 26; Bill C-27 2022, 30). Section 63 states on if an organization "... has used an automated decision system to make a prediction, recommendation or decision about the individual that could have a significant impact on them, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision" (Bill C-11 2020, 27; Bill C-27 2022, 31).

Unlike the earlier Bill C-11, Bill C-27 proposes the enactment of a wholly new *Artificial Intelligence and Data Act* (AIDA). The stated purpose of the AIDA is “...to prohibit certain conduct in relation to artificial intelligence systems that may result in serious harm to individuals or harm their interests” (Bill C-27 2022, 87). Essentially, the AIDA aims to establish a framework for accountability regarding the development and deployment of AI systems. Additionally, the AIDA is a more robust regulatory tool than the *Directive*.

The first way the AIDA aims to establish the accountability framework is through definitions that provide scope for the act. The AIDA provides a more precise definition of artificial intelligence than the *Directive*. The *Directive* definition reads: “Information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviors, or solving problems” (Canada 2019b). In contrast, the AIDA definition reads: “artificial intelligence system means a technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions” (Bill C-27 2022, 86). Furthermore, unlike the *Directive*, the AIDA provides a definition of ‘biased output’ as:

content that is generated, or a decision, recommendation or prediction that is made, by an artificial intelligence system and that adversely differentiates, directly or indirectly and without justification, in relation to an individual on one or more of the prohibited grounds of discrimination set out in section 3 of the *Canadian Human Rights Act*, or on a combination of such prohibited grounds (Bill C-27 2022, 86).

However, confusingly, the definition goes on to state the definition of bias does not include “...content, or a decision, recommendation or prediction...” that is intended to prevent “...or to

eliminate or reduce disadvantages that are suffered by, any group of individuals when those disadvantages would be based on or related to the prohibited grounds (Bill C-27 2022, 87).

The AIDA then defines who is responsible for an AI system as “...person is responsible for an artificial intelligence system, including a high-impact system, if, in the course of international or interprovincial trade and commerce, they design, develop or make available for use the artificial intelligence system or manage its operation” (Bill C-27 2022, 8788). The term ‘high-impact system’ is not defined in the AITA, but it notes that it will be defined in future regulations (Bill C-27 2022, 87). Notably, this definition identifies designers, developers, and managers as responsible for the AI system. The person responsible for an AI system must “...establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system” (Bill C-27 2022, 89). The AITA does not define the measures that need to be taken, that is also left to future regulations.

The second way the AITA establishes an accountability framework for AI is through granting powers to the Minister of Innovation, Science, and Industry to audit and sanction individuals responsible for AI systems (Bill C-27 2022, 90). The audit can be conducted by the Minister or by an independent auditor engaged by the Minister (Bill C-27 2022, 90). Following the audit, the Minister will have the power to order the audited party to “...implement any measure specified in the order to address anything referred to in the audit report (Bill C-27 2022, 91). Additionally, the AITA grants the Minister the power to designate a senior official the Artificial Intelligence and Data Commissioner (Bill C-27 2022, 97). This Commissioner would have all the powers of the Minister in auditing and sanctioning individuals responsible for AI (Bill C-27 2022, 97).

## Section 4: Analysis of AI Governance in Canada

This section analyze and critique AI governance in Canada. The first part of the section will examine Treasury Board *Directive on Automated Decision Making* in the public sector. The second part of this section will analyze the AI governance that is proposed in Bill C-27. Both current and proposed governance approaches will be evaluated for how they address challenges associated with the use of AI systems, how they establish accountability, and, in the case of Bill C-27, how they compare to AI governance in the EU.

### Directive on Automated Decision Making

The first critique of the Canadian Directive on automated decision making is that it amounts to little more than a simple questionnaire (Selbst 2021, 139). Selbst argues that it is important for an impact assessment to ask open ended and flexible questions because they lead to more fruitful and instructive responses (2021, 148). The questions that Canada poses are either yes/no questions or drop-down menus with answers provided (Selbst 2021, 148). For example, the algorithmic impact assessments ask: “Will the Automated Decision System use personal information as input data? Who controls the data? (Canada “Algorithmic Impact Assessment” quoted in Selbst 2021, 148)

Selbst contrasts this example with an open-ended version of the question “Describe all the data sources you used to train the model, including where they came from, who controls them, why you chose to use those datasets, and what impacts you anticipate from those choices?” (Selbst 2021,149). Fundamentally for Selbst, top-down questions like the ones asked by Canada always produce a “coarse and general” picture compared to open ended questions (Selbst 2021, 149).

The shortcomings of yes/no questions are illustrated in the report from the May 2022 impact assessment for the *Advanced Analytics Triage of Overseas Temporary Resident Visa Application* undertaken the Department of Citizenship and Immigration. Question 7 under ‘De-Risking and Mitigation Measures – Data Quality’ asks: “Do you have documented processes in place to test datasets against biases, and other unexpected outcomes? This could include experience in applying frameworks, methods, guidelines, or other assessment tools?” (Canada 2022g). Contrast this to questions 16 and 17 in the ‘Impact Assessment’ section of the questionnaire. Questions 16 asks “How long will impacts from the decision last?” with the selected response from a drop-down menu “Some impacts may last a matter of months, but some lingering impacts may last longer” (Canada 2022g). Question 17 asks for a follow up with “Please describe why the impacts resulting from the decision are as per selected above” and a 149-word response was submitted that thoroughly explains the selection.

Presumably it is more time consuming to manually assign an impact assessment to an AI system than to have a questionnaire that automatically scores and assigns the impact level. Yet, as indicated earlier in this capstone, a search on the Open Government Portal shows that only five such algorithmic impact assessments have been completed as of July 2022.

Second, Canadian approach to impact assessments is not an explicitly mandated early-stage intervention (Selbst 2021, 146). The landing page for the questionnaire only states that the impact assessment should be completed “at the beginning of the design phase of a project” (Canada 2022a). Also note that the questionnaire explicitly allows respondents to advance into the questionnaire proper even if the respondent selects that the project is in the implementation phase on the landing page. One of the critical goals of impact assessments is that they consider social impact early on so that their findings can inform projects before they are completed (Selbst 2021,

147). Another critical goal of impact assessments is informing the public and evaluators of design decisions (Selbst 2021, 147). Because of the complexity of AI systems, and the complexity of the design process, "...trying to parse a causal explanation for some error is impossible, unless we have documentation about what choices were made and why..." (Selbst 2021,147).

A related point is that the questions contained in Canada's impact assessment questions only considers an AI system in isolation and not as part of sequence of AI systems. That is to say, the questionnaire does not identify whether an AI system that is individually 'low risk' may be contributing to a chain of decisions that collectively could be considered 'high risk'. Buolamwini shows that even though AI may not be used directly in a high stakes task such as determining the length of a prison sentence, AI can be used in the chain of decisions leading to a high-risk task (Buolamwini and Gebru 2018, 1). Buolamwini illustrates the point well in the following statement:

For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis. (Buolamwini and Gebru 2018, 1)

## Digital Charter Implementation Act, 2022

If it were proclaimed in its current form, Bill C-27, would be at least the beginning of effective governance of AI systems due to its robust accountability framework achieved through powers granted to the Minister. However, Bill C-27 has a long way to go before it is proclaimed and it is unlikely to be implemented in its current form. Even in its current form, Bill C-27 can still be critiqued for failing to contain provisions that address some challenges associated with the use of AI, and it can be critiqued in comparison to governance approaches taken in the EU.

First, Bill C-27, despite containing a definition of ‘biased data’, the document does not contain a provision that mandates the use of quality, non-biased data. As was discussed earlier in this capstone, when biased data is used to develop an AI system, that AI system can replicate the bias in its predictions. The EU proposal does mandate the use of “high quality data sets” to develop an AI system in order “...to minimize risks and discriminatory outcomes” (European Union 2022b).

Second, the ‘right to an explanation’ regarding decisions made by an AI system codified in Bill C-27 is not as robust as the ‘right to an explanation’ set out in the EU *General Data Protection Regulation* (GDPR). The GDPR explicitly states that a person “...shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” (European Union 2016, 46). Bill C-27 only states that an individual will have the decision by an AI system explained to them.

Lastly, the accountability regime proposed in Bill C-27 relies on audits performed after an AI system has been implemented. The reliance on audits shares shortcomings with the *Directive’s* after the fact algorithmic impact assessments. Audits completed after the fact do not discourage a project from committing harms during the design phase. Furthermore, also like after the fact impact assessments, it will be similarly challenging to divine a causal relationship in an audit of an AI project after the fact.

## Conclusion

This capstone provided an overview of the complex challenges that result from AI's technical and social dimensions. On one hand, AI systems are technical pieces of software, fed instructions, and asked to complete certain tasks. On the other hand, AI systems are technical and social systems that reflect the implicit values of the people designing the system, and biases in data.

It was shown that the social-technical nature of AI creates several governance challenges that policy makers need to keep in mind while considering options for governing AI. Through skewed outputs, AI systems can reflect the biases of their designers or replicate and magnify the biases contained in the data used to develop the systems. It can also be almost impossible to evaluate and explain the output of an AI system because of the highly complex techniques used to develop the system. This problem is compounded by software firms that go to great lengths to keep the source code for the AI systems they develop private and proprietary to themselves.

This capstone then explored three areas of literature on AI governance. It first surveyed perspectives centered on using principles and ethics documents as guidelines for AI development. It showed that the various principle documents create confusion and lack legitimacy on an international scale. From there, perspectives focused on using human rights as governance guide for AI were reviewed. Following this, the capstone showed how impact assessments and particularly human rights impact assessment operationalize human rights into a governance tool. Particular attention was paid to the importance of accountability frameworks in impact assessments.

The capstone then proceeded through two case studies on AI governance. The first case examined the European Union's approach to AI governance. Two European regulatory

instruments were examined and their strengths relative to Canada explored: the 2021 European Commission's *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*, and the *General Data Protection Regulation (GDPR)*. The second case was Canada proper and again two regulatory instruments were explored: the 2020 *Canadian Directive on Automated Decision Making* issued by the Treasury Board of Canada Secretariat, and Bill C-27, the *Digital Charter Implementation Act, 2022*.

An analysis of Canada's two regulatory instruments identified several weaknesses that could be improved. For the Treasury Board Directive, the questionnaire component of the algorithmic impact assessment features many yes/no and drop-down menu selectable responses compared to short answer questions, and it is not explicitly mandated that the questionnaire be completed during the design phase of the project. Bill C-27 could learn from the EU Proposal and GDPR by mandating the use of quality, non-biased data, and by setting out a right for a citizen not to be subject to a decision arrived at solely by an AI system.

AI and its challenges are in Canada. This capstone made recommendations to help Canada address those challenges without stifling AI's capacity to benefit society. The capstone recommended modest changes to strengthen the Directive on the use of AI in the public sector based on lessons from the literature on AI governance, and improvements to Bill C-27 based on an examination of the European Union's approach to regulating AI. These recommendations will strengthen AI governance and protect citizens from the harmful effects of using these tools.

## Policy Recommendations

This section provides recommendations to improve AI governance in Canada. Recommendations on enhancing AI governance in Treasury Board of Canada *Directive* will be supplied first. This will be followed by recommendations on changes to Bill C-27 to enhance its AI governance. With targeted improvements to the Treasury Board *Directive*, and to Bill C-27, Canada could be a global leader in AI governance.

### Directive on Automated Decision Making

As shown earlier in this capstone, the Treasury Board *Directive* is lacking in two areas. First, the questionnaire component of the algorithmic impact assessment features many yes/no and drop-down menu selectable responses compared to short answer questions. This results in less detailed responses and only a general picture of the impact associated with a given AI system. Second, not explicitly requiring the completion of the algorithmic impact assessment in the design phase of the project means that the impact assessments can be completed after project completion. This means that the main benefit of an impact assessment, to identify an impact and adjust the project accordingly during the design phase, is missed.

#### **Recommendation 1:**

A) In the Algorithmic Impact Assessment questionnaire:

- Convert a portion of yes/no and drop-down menu questions into open-ended long answer questions.

- Add a section to the questionnaire that explicitly requests information on how the proposed AI system fits into larger workflow structure and how it may be part of a sequence of decisions.

B) Strengthen the mandate that the Algorithmic Impact Assessment be completed in the design phase of the project.

### Bill C-27 Digital Charter Implementation Act, 2022

Bill C-27, the *Digital Charter Implementation Act*, was shown to be lacking in comparison to the EU proposal and the GDPR in two ways. First, unlike the EU proposal, Bill C-27 does not contain provisions that mandate the use of quality, non-biased data. Second, unlike the GDPR, Bill C-27 does not set out a right for a citizen not to be subject to a decision based solely on a decision by an AI system. The recommendations below correct those deficiencies.

#### **Recommendation 2:**

A) Follow the EU *Proposal* and include a provision that explicitly mandates the use of quality, non-biased data.

B) Follow the GDPR model of a ‘right to an explanation’ by including language that explicitly accords citizens the right to not be subject to a decision wholly made by an AI system.

## References

- Ananny, M., K. Crawford. 2018. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *New media & society* 20 (3): 973-989.
- Bill C-11, *Digital Charter Implementation Act, 2020*, 2<sup>nd</sup> sess., 43 Parliament. 2020. <https://www.parl.ca/DocumentViewer/en/43-2/bill/C-11/first-reading>
- Bill C-27, *Digital Charter Implementation Act, 2022*, 1<sup>st</sup> sess., 44 Parliament. 2022. <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading>
- Bolukbasi T., K-W. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding. Paper presented at 30<sup>th</sup> Conference on Neural Information Processing Systems (NIPS 2016). <https://papers.nips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf>
- Bovens, M. 2007. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13 (4): 447-468.
- Buolamwini, J., T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings on Machine Learning Research* 81: 1-15.
- Canada. Digital Government. 2022a. "Algorithmic Impact Assessment tool." Accessed July 10 2022. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#toc3-1>
- Canada. Open Government. 2022b. "Algorithmic Impact Assessment." Accessed July 18 2022. <https://open.canada.ca/aia-eia-js/?lang=en>
- Canada. Open Government. Algorithmic Impact Assessment Results. 2022c. "Advanced Analytics Triage of Overseas Temporary Resident Visa Applications." Accessed July 10 2022. <https://opencanada.blob.core.windows.net/opengovprod/resources/9f4dea84-e7ca-47ae-8b14-0fe2becfe6db/trv-aia-en-may-2022.pdf?sr=b&sp=r&sig=LjsTgLB7OhtHsZLCG22IW%2B%2BkFPQrmPUASLt1uVg9FtY%3D&sv=2015-07-08&se=2022-07-20T03%3A06%3A33Z>
- Canada. Open Government. 2022d. "Open Government Portal." Accessed July 10 2022. [https://search.open.canada.ca/en/od/?sort=score%20desc&page=1&search\\_text=AIA](https://search.open.canada.ca/en/od/?sort=score%20desc&page=1&search_text=AIA)
- Canada. Open Parliament. 2022e. "Bill C-11(Historical)." Accessed May 18 2022. <https://openparliament.ca/bills/43-2/C-11/>
- Canada. Digital Government. 2022f. "Algorithmic Impact Assessment tool." Accessed July 10 2022. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#toc3-1>
- Canada. Public Services and Procurement Canada. 2018. *Artificial Intelligence (AI) pilot project for surveillance of suicide-related related behaviours using social media. (1000196416)*. Accessed July 10, 2022. <https://buyandsell.gc.ca/procurement-data/tender-notice/PW-17-00809483>

- Canada. Treasury Board of Canada Secretariat. 2019a. *Directive on Automated Decision-Making*. Ottawa, Date modified 2021-04-01.
- Canada. Treasury Board. 2019b. *Policy on Service and Digital*. Ottawa, Date modified 2019-08-02.
- Canada. Treasury Board of Canada Secretariat. 2022g. *Algorithmic Impact Assessment Results: Advanced Analytics Triage of Overseas Temporary Resident Visa Applications*. <https://open.canada.ca/data/en/dataset/6cba99b1-ea2c-4f8a-b954-3843ecd3a7f0>
- Casey, B., A. Farhangi, and R. Vogl. 2019. "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise." *Berkeley Technology Law Journal* 34 (1): 143-188.
- CBinsights. 2016. "Artificial Intelligence Explodes: New Deal Activity Record for AI Startups." Accessed May 2, 2022. <https://www.cbinsights.com/research/artificial-intelligence-funding-trends/>
- Chouldechova, A. 2016. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *arXiv.org*. <https://arxiv.org/abs/1610.07524>
- Citizen Lab. 2020. *To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada* by Kate Robertson, Cynthia Khoo, and Yolanda Song. Toronto: University of Toronto.
- Coren, M.J. 2017. "It took (only) six years for bots to start ditching outdated gender stereotypes." Accessed May 15, 2022. <https://qz.com/1033587/it-took-only-six-years-for-bots-to-start-ditching-outdated-gender-stereotypes/>
- Danks, D., and London, A. J. 2017. "Algorithmic Bias in Autonomous Systems." Paper presented at the 26th International Joint Conference on Artificial Intelligence (IJCAI). <https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf>
- Devendorf, L., and E Goodman. 2014. "The Algorithm Multiple, the Algorithm Material". Conference presentation at Contours of Algorithmic Life. UCDavis, May 2014.
- Donahoe, E., M.M. Metzger. 2019. "Artificial Intelligence and Human Rights." *Journal of Democracy* 30(2): 115-126.
- Dourish, P. 2016. "Algorithms and their others: Algorithmic culture in context." *Big Data & Society* 3(2): 1–11.
- European Union. 2016. *EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018*. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>

- European Union. Council of the European Union. Permanent Representatives Committee. 2019. *Artificial intelligence b) Conclusions on the coordinated plan on artificial intelligence - Adoption*. Brussels.
- European Union. European Commission. 2018. *Communication From the Commission To The European Parliament, The European Council, The Council, The European Economic And Social Committee, And The Committee Of The Regions: Artificial Intelligence for Europe*. Brussels.
- European Union. European Commission. 2020. *White Paper: On Artificial Intelligence - A European approach to excellence and trust*. Brussels.
- European Union. European Commission. 2021. *Proposal for a Regulation Of The European Parliament And Of The Council: Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. Brussels.
- European Union. European Commission. 2022. “First regulatory sandbox on Artificial Intelligence presented.” Accessed July 10 2022. <https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented>
- European Union. European Parliament. Legal Affairs. 2020. *Framework of ethical aspects of artificial intelligence, robotics and related technologies*. Brussels.
- European Union. European Commission. 2022b. “Regulatory framework proposal on artificial intelligence”. Accessed July 7 2022. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Fjeld, J., N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. 2020. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.” *Berkman Klein Center for Internet & Society*.
- Floridi, L., and J. Cowls. 2019. “A Unified Framework of Five Principles for AI in Society.” *Harvard Data Science Review* (1.1).
- Geburu, T., J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64, no.12: 86-92. <https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext>
- Goffey, A. 2008. “Algorithm.” In *Software Studies: a lexicon*, edited by Mathew Fuller, 15-20. MIT Press: Cambridge.
- Guardian, the. 2016. “A beauty contest was judged by AI and the robots didn’t like dark skin.” Accessed May 10, 2022. <https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people>
- Hiller, Kaitlyn. 2021. Predictive Policing and the Charter. *Manitoba Law Journal* 224 <<https://canlii.ca/t/tssw>>, retrieved on 2022-03-01

- Horvitz, Eric. 2017. "AI, people, and society." *Science* (Vol 357, NO. 6346). <https://www.science.org/doi/10.1126/science.aao2466>
- International Association for Impact Assessment. 2022. "IAIA: The leading global network on impact assessment." Accessed June 10, 2022. <https://www.iaia.org>
- Jobin, A., Lenca, M. & Vayena, E. 2019. "The global landscape of AI ethics guidelines." *Nat Mach Intell* 1, 389–399 (2019). <https://doi.org/10.1038/s42256-019-0088-2>
- Kaminski, M.E. 2019. "The Right to Explanation, Explained." *Berkeley Technology Law Journal* 34 (1):189-218.
- Karlin, M., and N. Corriveau. 2018. "The Government of Canada's Algorithmic Impact Assessment: Take Two." Accessed June 6 2022. <https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f>
- Latonero, M. 2020. "AI Principle Proliferation as a Crisis of Legitimacy." *Harvard Kennedy School: CARR Center for Human Rights Policy* Issue 2020 – 011.
- Lemelin-Bellerose, Sarah. 2019. "Artificial Intelligence: Current Situation, Risks and Outlook." Canada. Library of Parliament. [https://lop.parl.ca/sites/PublicWebsite/default/en\\_CA/ResearchPublications/201906E](https://lop.parl.ca/sites/PublicWebsite/default/en_CA/ResearchPublications/201906E)
- Machine Learning Mastery. 2019. "What Are Word Embeddings for Text?". Accessed May 10, 2022. <https://machinelearningmastery.com/what-are-word-embeddings/>
- Malli Nisa, Melinda Jacobs, and Sarah Villeneuve. 2018. "Intro to AI for Policymakers: Understanding the shift." Prepared for Brookfield institute for innovation + entrepreneurship. Accessed April 4, 2022. <https://brookfieldinstitute.ca/intro-to-ai-for-policymakers/>
- Mantelero, A., M.S. Esposito. 2021. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." *Computer Law & Security Review* 41.
- Metcalf J., E. Moss, E.A. Watkins, R. Singh, and M.C. Elish. 2021. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts." SSRN. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3736261](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736261)
- Moss E., E.A. Watkins, R. Singh, M.C. Elish, and J. Metcalf. 2021. "Assembling Accountability: Algorithmic Impact Assessment For The Public Interest." Report prepared for *Data & Society* June 2021. <https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/>
- ProPublica.org. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." Accessed April 3 2022. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

- Seaver, Nick. 2017. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." *Big Data & Society* 4 (2): 1-12.
- Selbst, A.D., 2021. "An Institutional View of Algorithmic Impact Assessments." *Harvard Journal of Law & Technology* 35 (1): 117- 191.
- Smith, Mathew L. Smith., and Sujaya Neupane. 2018. "Artificial intelligence and human development: Toward a research agenda." White paper.
- Whittlestone, Jess., Jack Clark. 2021. "Why And How Governments Should Monitor AI Development." *arXiv* August 2021. arXiv:2108.12427v2 [cs.CY] 31 Aug 2021
- Wieringa, M. 2020. "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability." Conference presentation at *Conference on Fairness, Accountability, and Transparency* January 27–30, 2020, Barcelona, Spain.