

FIRST NATIONS AND ARTIFICIAL INTELLIGENCE



**Research and Data
Management Sector**



Author

Dr. Benjamin Wald

Contents

| | |
|--|----|
| Acknowledgments: | 3 |
| Executive Summary | 4 |
| 1. Introduction | 6 |
| 2. Overview of AI Technology | 8 |
| 2.1 Introduction | 8 |
| 2.2 A brief history of AI..... | 8 |
| 2.3 Symbolic AI | 11 |
| 2.4 Machine Learning | 12 |
| 2.4.1 Training Data | 15 |
| 2.4.2 Types of Machine Learning..... | 16 |
| 2.5 Who owns AI..... | 20 |
| 3. AI Legislation and Regulation..... | 22 |
| 3.1 Introduction | 22 |
| 3.2 The Artificial Intelligence and Data Act (AIDA)..... | 23 |
| 3.3 The UK's regulatory approach to AI safety: | 28 |
| 3.4 United States guidance on AI..... | 29 |
| 3.5 European Union Artificial Intelligence Act | 32 |
| 3.6 Comparing AIDA to international approaches..... | 34 |
| 4. Discrimination by AI..... | 36 |
| 4.1 Introduction | 36 |
| 4.2 Causes of Algorithmic Bias | 37 |
| 4.2.1 Bias from training data..... | 37 |
| 4.2.2 Bias from algorithm design | 38 |
| 4.3 Risk Factors for Algorithmic Bias | 39 |
| 4.3.1 Black-Box AI | 39 |
| 4.3.2 AI Designed Without Diverse Input..... | 40 |
| 4.3.3 AI Without a Human in the Loop..... | 41 |
| 4.4 Mitigating the Risk of Algorithmic Bias | 41 |
| 4.4.1 Auditing of AI | 41 |

| | |
|--|----|
| 4.4.2 Consultation with First Nations | 42 |
| 4.4.3 Right to Object to AI Decisions | 43 |
| 5. AI and Data Sovereignty | 44 |
| 5.1 Introduction | 44 |
| 5.2 First Nation data sovereignty | 45 |
| 5.3 AI training data and data sovereignty | 46 |
| 5.4 Generative AI outputs and data sovereignty | 49 |
| 5.5 AI identifying First Nations | 50 |
| 6. Opportunities for First Nations in AI..... | 52 |
| 6.1 Introduction | 52 |
| 6.2 First Nations Language Revitalization..... | 52 |
| 6.3 Support for First Nation Governments..... | 53 |
| 6.4 Preserving the Land | 53 |
| Glossary | 55 |
| References | 57 |

Acknowledgments:

The author would like to thank the Ministry of Health, who funded the development of this report. The author would also like to thank the people who shared their knowledge and expertise to help make this report possible. The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Executive Summary

Artificial Intelligence is a powerful and disruptive technology. It has a great deal of potential, but this potential comes paired with serious risks for First Nations. This paper lays out the nature of AI technology, the proposed regulations that could govern it, and the risks and potential benefits to First Nations that AI presents.

AI is the name for a number of different technologies, united by the fact that they replicate abilities traditionally thought to require human intelligence. The two main branches of AI are symbolic AI and machine learning. Symbolic AI is created by programming a computer with a series of symbols along with rules and axioms for the manipulation of these symbols. In this form of AI, the programmers need to identify the relevant rules for the manipulation of data. A prominent example of symbolic AI is the Deep Blue chess-playing program that famously defeated world champion Garry Kasparov in a six-game series in 1997. It worked by going through a huge number of possible chess moves, as many as 200 million a second, and evaluating them based on a number of programmed in criterion, such as square control, king safety, development, and so on. It would then select the best move according to these preprogrammed criteria.

Machine learning is the other major branch of AI, and it has been responsible for much of the meteoric progress in AI over the past few decades. Unlike symbolic AI, machine learning systems do not rely on human generated rules that are programmed in ahead of time. Instead, machine learning systems are capable of generating their own sets of rules from training data they are provided. To do this, they require potentially vast sets of training data from which to learn. For example, GPT-3 was trained using materials from books, articles, and websites totalling 300 billion words.

The AI and Data Act, or AIDA, is a bill that if passed would regulate AI in Canada. It would apply only to high-risk systems. The original law left the identification of high-risk systems to future regulations. A proposed amendment would lay out 7 categories of use, including hiring, provision of services, and biometric identification that would qualify as high-risk, with the Governor in Council able to modify, add to, or delete categories via future regulations. The Assembly of First Nations has submitted a brief to the Parliamentary Standing Committee on Industry and Technology in October of 2023 that is sharply critical of the bill, both in the process by which it has been crafted and the substance. On the process, they object that the government did not engage in Nation-to-Nation consultation with First Nations during the drafting of the bill. On the substance, the brief objects to the lack of independent enforcement in the bill, with the enforcement being overseen by the Minister of Industry who is also responsible for the act itself, and for growing the AI industry in Canada. This creates a conflict of interests in the enforcement of the bill. In addition, it objects to the exemptions contained within the act, with the act not applying to many key government branches. Finally, they object that the act is overly individualistic, excluding any protection from community harms or infringements of community rights.

AI might seem neutral, but it has been shown in many cases to be discriminatory. This is often referred to as algorithmic bias. Algorithmic bias can result from omissions or bias in the training data. If First Nations are not included in the training data, then AI systems may not work well for First Nations people. This can be extremely harmful, for example if AI is used in healthcare and results in misdiagnosis or other medical errors. The training data can also reflect societal biases, which can then be taken up into the AI system. As an example, Amazon created an AI algorithm to assess the CVs of job applicants and trained it on their past hiring practices. However, since more men than women were hired by Amazon, especially in tech roles, this resulted in an AI hiring system that was biased against women, downgrading resumes that contained the word “women’s” (as in women’s chess team) and graduates from two all-women’s colleges. The design of algorithms can also create bias, especially if a proxy is used to substitute for the true goal of the algorithm.¹ This choice of a proxy can itself embody bias if a proxy is chosen that does not track the actual goal in the case of First Nations peoples.

There are a number of things that can be done to help mitigate the risk of algorithmic discrimination. One is careful auditing of AI systems, both before they are introduced and over their lifecycle, to look for evidence of bias. Another is to ensure that First Nations are consulted during all stages of AI development, including the decision whether to use an AI system at all. And finally, a right to object to AI decision making can help in identifying and counteracting biased outcomes.

AI also raises important data sovereignty concerns. Data sovereignty is the right of First Nations to manage and control data about themselves, their culture, and their lands and resources. First Nation data being used in training machine learning systems without permission violates this right to data sovereignty. Additionally, generative AI violates data sovereignty when it creates outputs that mimic First Nation artwork and culture. Finally, AI can potentially identify First Nations communities in datasets that do not explicitly label First Nations. This expands the range of datasets that could reveal information on First Nations, and thus First Nations must assert their data sovereignty rights to a wider range of datasets.

Alongside these risks, there are potential benefits to AI technology for First Nations. One is First Nation language revitalization. The First Languages AI Reality, or FLAIR, project for example is seeking to create AI systems that can speak and understand First Nations languages. This could be used to help teach new speakers of these languages. A second opportunity comes from the potential for AI to help ease the administrative burden on First Nations governments, supporting First Nation self-governance. Finally, there are some heartening examples of AI being used by and with Indigenous peoples to help them preserve and revitalize their land that has been affected by climate change or other forms of damage.

¹ A proxy is a variable that is meant to track the desired outcome and is used when the desired outcome cannot be measured directly. For example, scores on standardized tests might be used as a proxy for educational success.

1. Introduction

Artificial intelligence, or AI, has the potential to be among the most important and disruptive technologies ever developed. It is a “general purpose technology”; a technology that has a wide range of applications across different industries and sectors. Already AI technologies have become ubiquitous, from the recommender algorithms that determine what posts we see on social media, to the voice recognition technology that powers digital assistants like Siri and Alexa, to behind-the-scenes uses of AI technology to assist companies in adjusting prices shown to different customers. This usage of AI is likely to expand in coming years, as the technology becomes more powerful and new systems are developed for different uses. The rate at which AI research has been developing over the last few years has been shocking, with AI systems growing in their capacities by leaps and bounds. This has been matched by the huge inflows of money into AI research. According to the Organization for Economic Cooperation and Development (OECD), around \$75 billion dollars of Venture Capital investment went into AI in 2020. That makes up 21% of the global Venture Capital investment in 2020. This is up dramatically from a 4% share, or around \$3 billion, in 2012.²

Alongside this potential, we also need to recognize that AI technology poses significant risks. Many of these risks apply only, or apply to a greater extent, to historically marginalized groups. There are numerous examples of AI systems that have produced biased and discriminatory results. From facial recognition technologies that fail significantly more often for women of colour,³ to AI-powered hiring algorithms⁴ that down-rank female applicants,⁵ to AI used by the criminal justice system to assess risk of reoffending that was biased against Black people,⁶ the evidence of AI discrimination is widespread. First Nations face these same risks of discrimination from the increased deployment of artificial intelligence, along with some further risks that are specific to First Nations.

It is increasingly important to be aware of what artificial intelligence is, how it is regulated, and how it might affect First Nations. This paper is intended for First Nations leadership and anyone else with a stake in how AI will affect First Nations, such as policy makers. It provides an introduction to AI technology, the ethical issues that arise for First Nations from this technology, and the opportunities it presents for First Nations. The discussion presupposes no prior knowledge of AI or AI ethics.

Section two of the paper provides a high-level overview of what AI is and the different varieties of this technology currently in use. It also provides a brief overview of the

² Amdur, 2023

³ Buolanwini and Gebru, 2018

⁴ An algorithm is a set of rules for a computer program to follow to arrive at an output.

⁵ Dastin, 2018

⁶ Angwin et al., 2016

history of AI technology, and who owns AI technology—including the training data that is used to build AI systems, the algorithm itself, and the outputs of the system.

Section three of the paper looks at the legislation and other forms of regulation that have been proposed for governing AI technology, both in Canada and abroad. It looks in detail at the proposed Artificial Intelligence and Data Act (AIDA) that is currently being debated in Canada and compares this act to the approach takes in the UK, U.S., and European Union.

Section four looks at the problem of discrimination by AI, often called algorithmic bias. This form of bias can easily affect First Nations and can produce a number of harmful and discriminatory results. The section breaks down the causes of this algorithmic bias, the risk factors that can produce this bias, and some ways to mitigate the risk of these forms of bias.

Section five considers the impacts of AI technology on data sovereignty. It provides a brief overview of what First Nations data sovereignty is. The section then looks at three risks to data sovereignty that arise from AI systems. The first is the risks to data sovereignty that are posed by the use of First Nations data to train AI systems. The second is the data sovereignty issues raised by the outputs of generative AI models, such as chatGPT or DALL-E. The third is the risk to data sovereignty raised by the ability of AI systems to identify First Nations groups even in data that has seemingly been stripped of First Nations identifiers.

Finally, section six looks at opportunities for First Nations arising from AI. It discusses the potential for AI to help with First Nations language revitalization, and important work already being done in this area. It looks at how AI could potentially provide support for First Nations governance, by easing administrative burdens. And finally, it discusses how AI systems can be used to help protect the land and aid First Nations peoples in keeping their lands healthy and flourishing.

AI technology is becoming an increasingly influential part of the technological landscape. It is important that First Nations leadership be able to advocate for AI that is fair, and that respects their data sovereignty rights, and that can provide benefits for First Nations. This paper aims to provide the tools and knowledge to support this advocacy.

2. Overview of AI Technology

2.1 Introduction

Artificial intelligence, or AI, is not a term for a single technology. Instead, AI is used to refer to a group of technologies that have similar effects. There have been many different definitions offered for AI. The Oxford dictionary defines artificial intelligence as “The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”⁷ Meanwhile, the Organization for Economic Co-operation and Development (OECD) defines artificial intelligence as “a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.”⁸ Both of these definitions will include a broad range of different technologies. This chapter will provide an overview of different technologies in the space of artificial intelligence.

For the purposes of this paper, we will be focused on technologies that currently exist or that might exist in the near future. For this reason, we will not be discussing so-called strong AI, also sometimes called artificial general intelligence. These would be AI systems that are capable of doing everything a human reasoner can do, and potentially more. Instead, we will focus on weak AI, or artificial narrow intelligence. This refers to AI systems that are able to do tasks that traditionally have required human intelligence, but only for specific narrow areas of application.

The two major types of AI systems we will look at are symbolic AI and machine learning. Symbolic AI, sometimes referred to as good old-fashioned AI, abbreviated to the acronym GOF AI, was the dominant approach to AI from the 1950's to the late 1980's. It refers to AI that works by explicitly representing human knowledge in a symbolic form. Machine learning, on the other hand, are systems that are able to learn and adapt without following explicit instructions by identifying patterns in the data. Machine learning is responsible for much of the modern expansion in the use of AI, leveraging huge datasets to come to novel and surprising conclusions. Both of these forms of AI will be unpacked in more detail below.

2.2 A brief history of AI

The history of Artificial Intelligence goes all the way back to the invention of the digital computer in the 1940's. In the year 1950 Alan Turing, who was one of the originators of modern computer science, wrote a paper exploring the possibility that a machine could think.⁹ At first, research was hampered by the physical restrictions of early digital

⁷ Oxford Dictionary of Phrase and Fable (2 ed.), 2005

⁸ OECD, 2019

⁹ Turing, 1950

computers, and by the high cost of computation. Still, research on artificial intelligence continued throughout the 1950's and 1960's, with the concepts of machine learning and artificial neural networks, two of the main pillars of current AI systems, being formalized during this time.¹⁰ However, the dominant approach between the 1950's and 1970's was symbolic artificial intelligence.

Major strides in AI were made during this period, including the creation of a mechanical mouse that could navigate a maze, and a program that could play checkers well enough to challenge a beginning player. Proponents of AI made grandiose claims, with several pioneers in the field claiming that AI would match the intelligence of humans within 10 years.¹¹ In part because of these overhyped claims, the field suffered a major slowdown between the 1970's and the 1990's. This period is often referred to as the "AI winter", as funding for artificial intelligence dried up and interest receded.¹²

The early 1980's saw the rise of commercial "expert systems", symbolic AI systems that applied rules-based procedures to a body of pre-specified knowledge to arrive at outputs. These systems were enormously popular, but they were explicitly domain specific and did not seek to emulate intelligence on a general level.¹³

The 1990's saw a resurgence of interest in machine learning techniques, rather than the traditional symbolic artificial intelligences.¹⁴ The decade also saw a landmark achievement in AI, with IBM's Deep Blue defeating reigning chess champion at the time Gary Kasparov. This was the first program to defeat a human chess champion. Deep Blue used traditional symbolic AI and did not incorporate the machine learning techniques that were to become so central to modern AI.

The 2000's were when machine learning began to hit its stride, making major progress in speech and image recognition tasks. Many major commercial uses of these technologies began to appear during this time, such as the release of Apple's virtual assistant Siri in 2011, or the introduction of a recommender algorithm into Facebook in 2009 to rank posts, as opposed to simply displaying them in the order they were posted. While recommender algorithms existed in the 1990's, they became much more advanced and widespread in the 2000's. 2011 also saw another milestone, with IBM's machine learning system Watson defeating Ken Jennings and Brad Rutter at Jeopardy.

The last decade has seen huge advances in AI technology, with the progress accelerating over time. Ten years ago, no AI system could reliably provide language or image recognition at a human level. However, today AI systems can beat humans in

¹⁰ Macukow, 2016

¹¹ Anyoha, 2017

¹² Schuchmann, 2019

¹³ Arif, 2023

¹⁴ Foote, 2021

tests in these areas.¹⁵ Language recognition has gone from a difficult and error-prone field to become ubiquitous, included in smart phones and dictation software that are a part of our everyday lives.

One of the most dramatic improvements over the last decade has been in generative AI. Generative AI is AI that can produce its own content based on prompts, rather than just categorizing inputs that are provided. In 2014, an AI generated face was a pixelated blur. By 2017, AIs could generate a crisp, clear image of a human face. In 2021, AI could not only generate a human image, but create an entire scene based on a short description, complete with background and surrounding details. And in 2024, generative AI can now create entire videos based on prompts. Text generation has similarly grown by leaps and bounds.

This rapid progress has been due to several factors. The main driver of these advances has been machine learning. Part of the reason machine learning has come into its own in the past decade is due to advances in machine learning techniques. In addition, the exponential growth in the power of computers has allowed AI researchers to scale up the complexity of their AI models.¹⁶ A further crucial component is the accessibility of massive quantities of data on which to train these AI models. In 2010, an estimated 2 zettabytes of data were generated globally. A zettabyte is equal to a trillion gigabytes of data. By 2020, this has increased to 64.2 zettabytes, and in 2023 this had grown to an estimated 120 zettabytes. This trend of ever-increasing data production is forecast to continue.¹⁷ This almost unimaginable quantity of data has been a key driver of the advancement of AI models, with each new model requiring more training data than the last.

¹⁵ Roser, 2022

¹⁶ Thompson et al., 2022

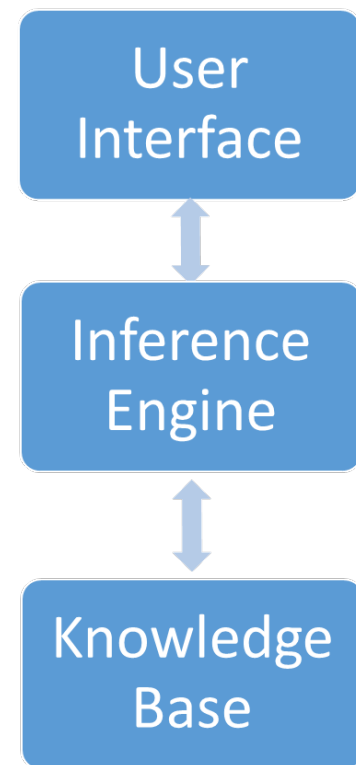
¹⁷ Duarte, 2023

2.3 Symbolic AI

Symbolic AI is created by programming a computer with a series of symbols along with rules and axioms for the manipulation of these symbols. For example, a medical expert system might take as an input a series of patient symptoms, each one represented by a particular symbol within the program and apply a set of programmed rules to these symbols to arrive at a diagnosis. The rules programmed into the computer might include one that states that if a patient has a cough, a fever, and difficulty breathing, then it should output a diagnosis of pneumonia.¹⁸ A prominent example of symbolic AI is the Deep Blue chess-playing program that famously defeated world champion Garry Kasparov in a six-game series in 1997.

Symbolic AI has a number of important strengths. Since it follows rules that have been programmed by humans, its operation is transparent. In other words, if we want to know why the system came to a particular output, we can trace this back to the set of rules that it was following to arrive there, and thus explain this output in terms of the design of the system. In addition, it does not need a large pool of data to be created. As long as the programmers can identify the correct rules for the system to follow, a symbolic AI can be created even in the absence of reliable data.

However, symbolic AI also has some significant limitations. It is in many ways limited by the extent of existing human knowledge. A symbolic AI relies on the programmers knowing the correct rules to program into the system for it to produce the desired output. This does not mean that humans would necessarily be able to produce the same answer as the AI; Deep Blue was able to beat the best human chess player after all. This is because Deep Blue was able to examine as many as 200 million chess positions per second, far exceeding human capacity. However, it still relies on clearly defined rules that can be programmed into a computer system and applied to a clear set of data—chess has clear rules and victory conditions that can serve as the basis for symbolic AI. This limits the applicability of symbolic AI in areas where programmers do not necessarily know the correct rules for the system to follow. A good example of this is machine vision; it has proved enormously difficult to come up with rules for identifying



This image demonstrates the flow of data in an expert system

¹⁸ Datacamp, 2023

even simple objects in visual images. As we shall see, this is an area of strength for machine learning.

Symbolic AI also struggles as the number of potential outcomes expands. As the number of symbols and rules increase, so do the computational demands on a symbolic AI. While symbolic AI was able to create Deep Blue to play chess, there was very limited success at the game of Go. This is because Go has dramatically more options for each move than chess. The number of legal board positions in Go has been calculated to be approximately 2.1×10^{170} , which is far greater than the number of atoms in the observable universe.¹⁹ This makes the computational resources required to apply symbolic rules to the game unfeasible.

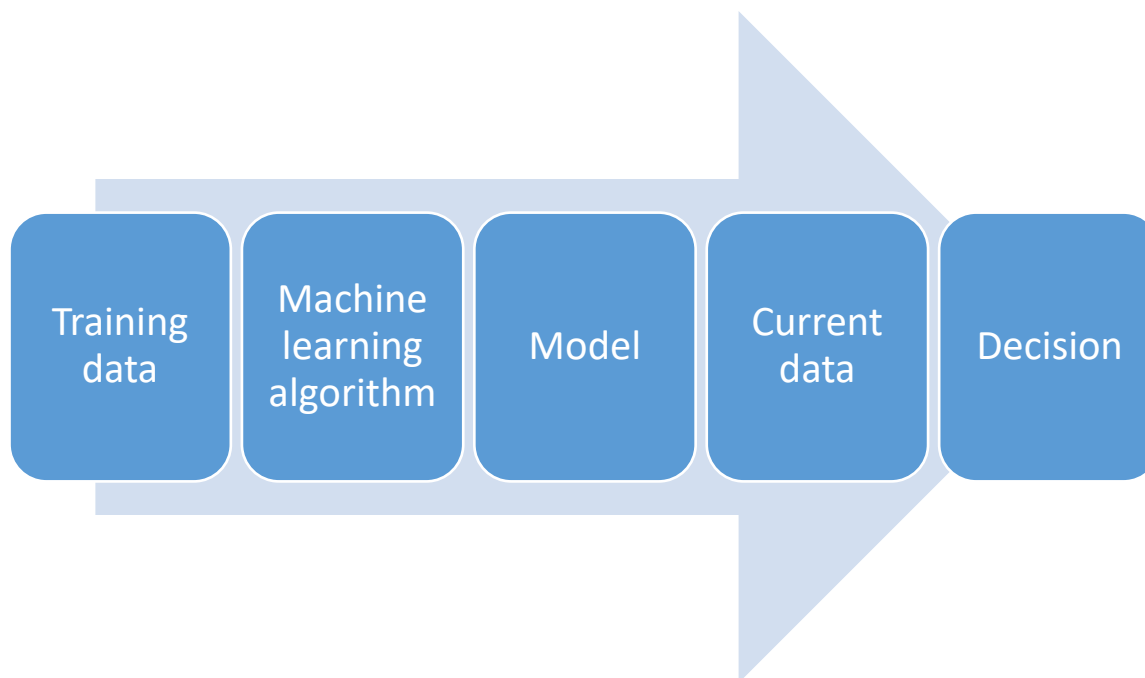
Symbolic AI also has no ability to adapt or learn over time. The only way to change the system is to manually reprogram it, and it will not learn from its own successes or failures along the way. Human beings would have to observe the systems outcomes and determine how best to change the system to improve on those outcomes.

2.4 Machine Learning

Machine learning is the other major branch of AI, and it has been responsible for much of the meteoric progress in AI over the past few decades. Unlike symbolic AI, machine learning systems do not rely on human generated rules that are programmed in ahead of time. Instead, machine learning systems are capable of generating their own sets of rules from training data they are provided.

The basic machine learning approach is to start with a set of training data for the desired task. The training data will be a pool of historical data that will be used to train the AI system. Programmers will then apply a machine learning algorithm to this training data. The machine learning algorithm in turn generates what is called a model—a set of rules that, when applied to the data, will provide an output. The model, like a symbolic AI system, will consist of a set of rules that transform data inputs into an output. The difference is in how this model is created—not by humans choosing a set of rules and programming them, but instead by the machine learning algorithms operating on the data.

¹⁹ Tromp, 2016



This diagram demonstrates the steps of developing a machine learning system

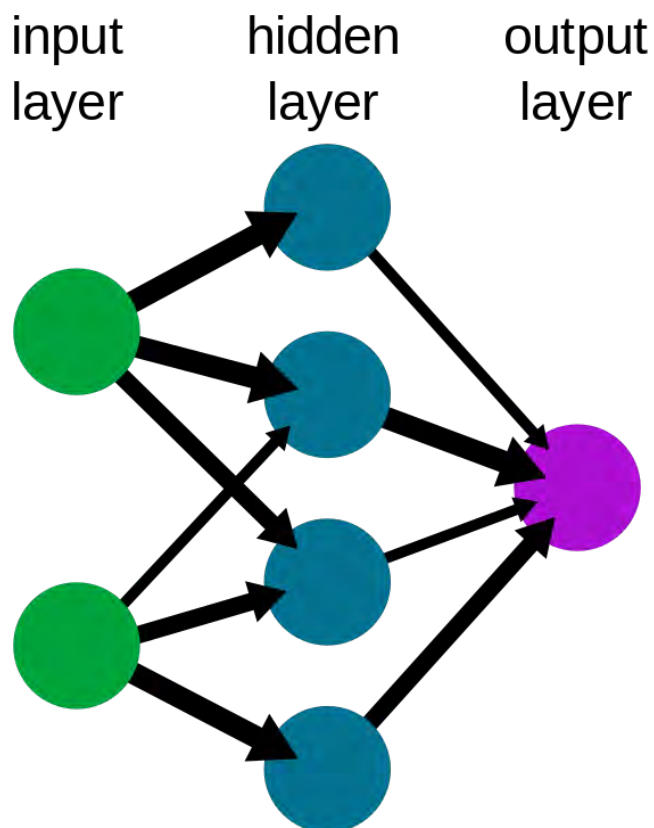
So how do machine learning algorithms create a model? The machine learning algorithm begins with a random guess—in other words, a random set of rules that will take the data inputs and provide an output. This will essentially always be bad at arriving at the desired output. The system will then begin to tweak the model in various ways, becoming more accurate over time. The machine learning algorithm will make a huge number of such tweaks, seeking to become more and more accurate. At the end of this process the machine learning algorithm will arrive at a final model, which can then be fed with current, rather than historical, data to arrive at a decision.

To make this more concrete, consider a specific example. Imagine that a company wants to develop a system to identify the perceived gender of an individual based on a picture of their face.²⁰ The first step would be to assemble a set of training data. In this case that would be a set of facial images, labelled with a gender tag. The machine learning algorithm would begin with a random model—a random set of rules to follow to map facial features on to perceived gender. This will be very inaccurate. The machine learning algorithm will then begin tweaking the model, by comparing the outputs of the model to the labels provided with the images. Different machine learning algorithms use different techniques, but the basic principle is to apply an algorithm that will slowly, over

²⁰ This is obviously problematic, since people's biology does not reliably track with their gender, and fails to consider people who are non-binary, two-spirited, trans, and so on. However, this is something that AI has been used for in the past and will provide a clear example.

many iterations, improve the performance of the model. Eventually, if the training is successful, the model will accurately be able to predict the labels provided for the faces in the training data. At this point, the model would be fed a new set of faces, not included in its training data. This is called the test data set. If it is sufficiently accurate on this new data set, which will mean providing a perceived gender that matches the labels attached to each image, then the model will be a success. This model is now finished training and can be put into use. It will be provided with new facial images that do not yet have a label and will be able to provide its own assessment of the perceived gender of these faces.

A simple neural network



One common variety of machine learning algorithm is artificial neural networks. This is a variety of machine learning that processes data in a way inspired by the human brain. Artificial neural networks are made up of a series of nodes, arranged in layers. The input is fed into the nodes of the input layer. The data is then processed through a number of so-called “hidden layers”. Each node will have connections to all of the nodes in the next layer, with the strength of these connections being what will be tweaked by the machine learning algorithm. The final results are fed to the output layer, which provides the AI’s final decision. This is somewhat akin to the way neurons in our brains work, with each neuron having different strengths of connection to

nearby neurons, and the strengths of these connections being shaped by experience over time.²¹ This is just one variety of machine learning algorithm, and there are many others with different strengths and drawbacks.

One of the results of machine learning is that it can be difficult or even impossible for the designers of a machine learning system to know how the system reaches the

²¹ Gurney, 2018

outputs that it does—these systems are so-called “black boxes”. This is because the model was built by the AI learning algorithms, rather than any human decision maker. The rules that a model is following can be enormously complex and can pick up on patterns in the data that humans aren’t able to discern. This is part of the strength of machine learning—it can learn to detect new and surprising patterns in data that humans cannot identify. However, this also leads to a lack of transparency in these systems. As an example, we can look at the machine learning system AlphaGo. Like Deep Blue, AlphaGo is an AI system designed to play a game, in this case the game of Go. However, unlike Deep Blue which relied on symbolic AI, AlphaGo was trained using machine learning techniques. As a result, some of its strategies and moves did not resemble any existing human strategies; it learned to play Go in its own unique way.²² It had great success; in 2017 AlphaGo beat Kie Jie, the number 1 ranked Go player in the world.

2.4.1 Training Data

Machine learning AI systems rely on huge quantities of training data. The accuracy of the machine learning system will depend on the quality of the data that it is provided with. There are a number of ways that a mismatch between the training data and the task the AI is being designed to perform will result in error, which is important to recognize. These sources of error can also become sources of bias, as will be discussed further in the section on discrimination by AI.

One prominent source of error from training data is unrepresentative data.

Unrepresentative data is data that differs in important ways from the real world. If the training data for an AI is not representative, the system may learn patterns that exist in the training data, but do not exist in the real world. A famous example is training an AI to distinguish between pictures of wolves and pictures of dogs. If it happens to be that all of the pictures of wolves include snow in the background, while the pictures of dogs do not, then the final AI system may learn this pattern instead of focusing on visual differences between wolves and dogs. The system will have learned to identify snow, rather than wolves. When deployed in the real world, this correlation will no longer hold, and the system will systematically misidentify wolves and dogs based on the presence or absence of snow.²³

Another potential source of error that can arise in training an AI is when the data is tracking a proxy variable for the desired outcome, rather than tracking that outcome directly. A proxy is a variable that is meant to track the desired outcome and is used when the desired outcome cannot be measured directly. Relying on a proxy may be unavoidable at times, but inaccuracy could result if the proxy selected does not always track the desired outcome. As an example, there was an AI system used in the United

²² Metz, 2016

²³ Ribeiro et al., 2016

States to assess patients' medical need, to assign additional clinical resources to high needs patients.²⁴ However, since medical need could not be measure directly, they used health care spending on individuals as a proxy variable. The idea was that those with more medical need would also incur more health care spending, and so this proxy would track medical need. However, given bias in the healthcare system, it turned out that Black patients tended to have fewer health care resources allocated to them, and hence lower health care spending, than white patients with equivalent medical need. Thus, the proxy variable failed to accurately track the intended target, and the AI system trained on this data ended up assigning Black patients lower levels of medical need than white patients. This is of major concern to First Nations, since First Nations are at risk being subject to this kind of discrimination from AI systems if they are not properly designed and regulated.

2.4.2 Types of Machine Learning

Machine learning can be divided up based on the type of training data that it uses.

There is supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning is learning that uses labelled training data. Thus, for each piece of training data, there is an associated label that corresponds to the variable that the AI system is trying to predict. If the system is trying to learn to predict which medical images show signs of cancer, the training data will consist of a large pool of medical images, which each image labelled as to whether or not it shows signs of cancer. The machine learning system is fed the raw images and tries to learn to predict the correct associated label. The accuracy of the system is assessed based on how well it can predict the labels provided for its training data.

A few things are important to note about supervised learning. It relies for its operation on the existence of huge, labelled datasets. While these may sometimes already exist, as perhaps would be the case for the medical images example given above, in many cases the raw data exists but the labels do not. In these cases, the data must first be labelled by humans before it can be used to train AI systems. Much of the work of labelling data is outsourced to poorly paid workers in the Global South.²⁵

The reliance on labels for the data introduces another potential source of error into AI systems. If the labels are inaccurate, or show signs of bias, then the AI system will learn to reproduce these same inaccuracies and biases. For example, an AI system might be designed to assess loan applications and recommend which ones to accept and which to reject. This might seem like a route to avoiding human biases in loan assessment. However, the training data might be a set of loan applications received in the past, and the labels provided by whether the human employees of the bank accepted or rejected the application. If this was the approach taken, then any biases in the way that

²⁴ Obermeyer et al., 2019

²⁵ Chandran et al., 2023

employees assess loan applications are likely to be learned by the AI system and reproduced.

As a concrete example of this kind of bias, we can look to Amazon's attempt to use AI in its hiring process. Amazon designed this AI system to assess candidate's resumes and give them a ranking of one to five stars. The data used to train the model was taken from resumes submitted to the company over a 10-year period. However, most of these applications came from men, due to the overrepresentation of men within the tech industry. As a result, the model developed a bias in favor of male candidates. 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, according to people familiar with the matter.²⁶ The program was subsequently scrapped, but this demonstrates how biases in training data can be replicated, rather than corrected, by machine learning tools.

Unsupervised learning, on the other hand, operates on unlabelled data. Rather than learning to predict specific labels, unsupervised learning systems learn to group data together according to patterns that are detected in the data. In this way, unsupervised learning can often discover patterns that humans would never have recognized. Unsupervised learning is used by things like AI recommender systems, which learn from huge volumes of purchases to recommend items to users based on what other similar users have bought in the past.

Unsupervised learning is not immune to inheriting human biases, however. The groupings in the data that unsupervised learning learns to predict can themselves be the product of bias. For example, Google's autocomplete feature uses unsupervised learning to predict how to complete a search query. However, in the past this has led to prejudiced and offensive autocomplete results, such as completing the query "are Jews" with suggestions such as "evil".²⁷ While this and many other such results have been manually corrected by Google once they have been made public, other biased autocomplete results continue to surface.

Reinforcement learning is the final major machine learning methodology. In reinforcement learning, AI systems are given a "reward function". This reward function specifies rewards for particular outcomes. The AI system then experiments with different actions and observes the reward that it receives. The system continues to try different actions, seeking over time to arrive at a policy that will maximize the amount of reward it receives. Unlike supervised learning, reinforcement learning does not require

²⁶ Dastin, 2018

²⁷ Cadwalladr, 2016

labelled training data. Instead, the system learns through trial and error what actions will arrive at its goals.

One of the major challenges for reinforcement learning is specifying a reward function. It is often quite difficult to lay out all the aspects of the desired outcome fully and completely. AI systems can end up engaging in what is called “reward hacking”, where the system maximizes its overall reward function, but does so in a way that fails to achieve the actual goals of the systems creators. As an example, reinforcement learning was used to train an AI system to play a boat racing game, where it was rewarded for reaching certain landmarks and for finishing the race. Rather than running the race as intended, the AI system learned to drive in circles past the landmarks, never actually completing the race, as this maximized its reward.²⁸

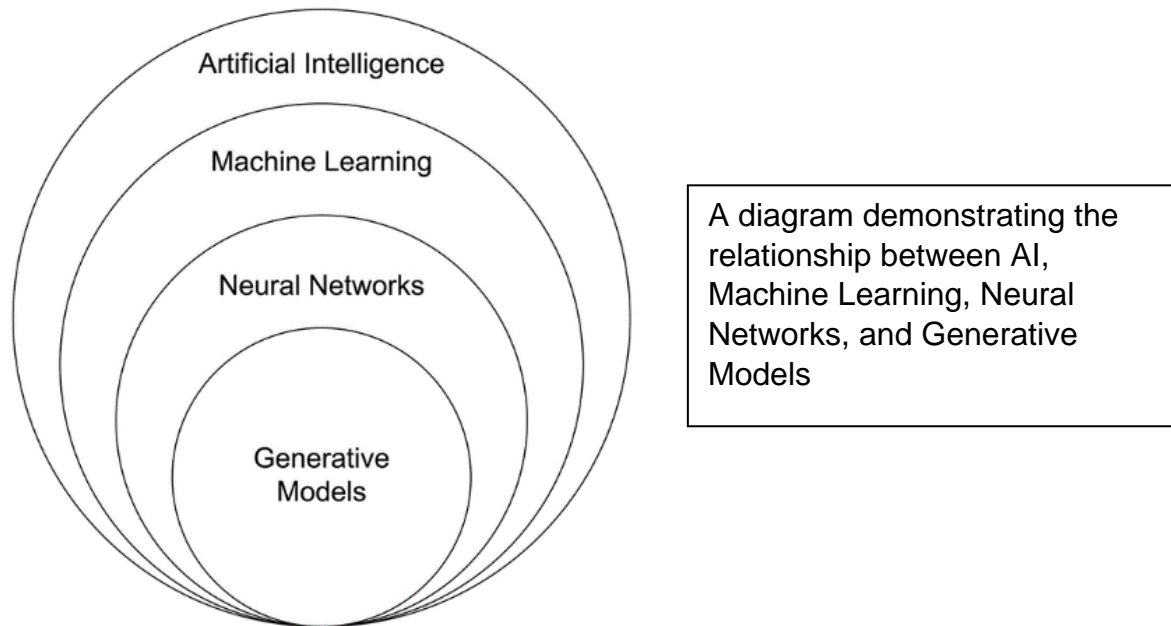
A more troubling example of reward hacking comes from large language models (LLMs), a form of generative AI (which will be discussed further below). Large language models produce text in response to a prompt; ChatGPT is probably the most famous example. These systems can be trained using reinforcement learning, which rewards them for producing stylistically appropriate content. However, this can lead to LLMs to invent plausible sounding but false content in response to prompts.²⁹ As an example, an LLM designed to help scientist write papers by summarizing research results began to invent fake scientific papers, sometimes even attributing them to real researchers.³⁰ This kind of plausible but false result can maximize the reward function, but clearly leads to very undesirable and potentially dangerous results.

One of the most talked about advances in AI has been the development of **generative AI**, such as ChatGPT and DALL-E. These are systems that go beyond classifying inputs into groups as traditional machine learning systems do and are instead capable of generating their own content. ChatGPT, for example, can produce text on a variety of topics and in different styles based on a prompt provided. This can be anything from writing fiction in the style of a particular author, to producing a credible university essay on a specified topic. DALL-E can produce novel artwork based on a provided prompt, showing a remarkable ability to mimic particular art styles and to include specific elements requested.

²⁸ Clark and Amodei, 2016

²⁹ Babu 2023

³⁰ Heaven, 2022



Generative AI is based on the same principles as other forms of AI. ChatGPT, for example, works in basically the same way as the autocomplete function on Gmail does. When you start typing a sentence in Gmail, it tries to predict the most likely next word in the sentence on the basis of its training data and offers this as a potential autocomplete option. ChatGPT does the same thing—on the basis of the prompt, it predicts the most likely next word, then the most likely word to come after that, and it continues this process until it has generated a full paragraph or more of text.

One of the differences between generative AI and other forms of AI is that generative AI tends to require even more training data. GPT-3 was trained on data obtained from books, web texts, Wikipedia, articles and other pieces of writing on the internet. In total, 300 billion words were used as the training data.³¹ Similarly, DALL-E was trained on a massive dataset containing millions of images and corresponding text descriptions, created by crawling the internet and collecting images from various sources, including social media, search engines, and image-hosting websites.³² In both cases, the data used for training the AI system contained multiple pieces of copyrighted work, which has led to lawsuits over both of these systems for copyright infringement. Courts have not yet determined the legality of training AIs on copyrighted works, so the outcomes of these cases stand to have major impacts on how generative AI is used and trained going forward. There are also significant First Nations data sovereignty issues with how generative AI is trained, since First Nations data was also used in the creation of these systems. This will be discussed in more detail in chapter 5.

³¹ Hughes, 2023

³² Pocock, 2023

2.5 Who owns AI

AI technology is worth a lot of money. According to the Organization for Economic Cooperation and Development (OECD), around \$75 billion dollars of Venture Capital investment went into AI in 2020. That makes up 21% of the global Venture Capital investment in 2020. This is up dramatically from a 4% share, or around \$3 billion, in 2012.³³ Of the biggest 10 companies by market cap, five of them are big tech companies that are heavily invested in AI technology: Microsoft, Apple, Amazon, Google, and Meta (Facebook). A sixth, NVIDIA, is on the list because they manufacture the computer chips used for AI systems, and as a result their stock has more than quadrupled in price over the last year, with sales expected to be up 60% in 2024³⁴.

With all of this money at stake, the question of who owns AI technology is a crucial one. There are a number of different answers to this question, depending on what exactly we mean. We can raise questions of ownership about three main components of AI systems. The first is ownership of the training data that is used to train a machine learning AI system. The second is ownership of the AI algorithm itself, which takes in data and outputs results. The third is the ownership of these outputs.

Consider the training data used to train a machine learning system designed to recognize faces. This training data will consist of a large number of images of faces. Imagine this data was gathered by scraping images from a popular social media site, such as Facebook.³⁵ Who owns this data? Does each individual image belong to the person whose face it is? Is it owned by the group or corporation that gathered the images into a single database? Or by the social media company who made the images available online in the first place? In fact, no one owns the data. Canadian law does not recognize any free-standing ownership in data.³⁶ While individuals can have a privacy interest in their data, and thus have certain rights and protections, they do not legally own it. In *McInerney v. McDonald* the Supreme Court considered and rejected the claim that individuals have property rights in their personal data.³⁷ Meanwhile, corporations can have patents on means of analysing data, or copyright over a particular database if sufficient skill and judgment go into its selection and arrangement, but they cannot own the underlying data itself. Thus, training data itself cannot be owned in the traditional sense, although there are various rights and obligations associated with how and when data can be used.

The algorithms themselves that are created and trained on this data can be owned. Specifically, these algorithms can be protected by patent law. Many of the basic techniques used in artificial intelligence were developed in academia and are free for anyone to use. Some AI systems have been made “open source” and are similarly

³³ Amdur, 2023

³⁴ Dyvik, 2023

³⁵ “Scraping” data refers to having a program go through a publicly available source of data and systematically extract this data. See glossary for a more thorough definition.

³⁶ Scassa, 2018

³⁷ *McInerney v. MacDonald*, 1992

freely available for others to use. However, many of the modern advances in artificial intelligence have been developed and perfected by private corporations and are protected by patents. These patents are the source of a great deal of wealth for the companies that control them.

When it comes to the outputs of AI models, in many cases this will be another form of data, and thus again cannot be owned. For example, if an AI system generates predictions about who is at risk for a particular disease, no one owns these predictions. However, the question becomes more complicated when we turn to generative AI and consider the outputs of such systems. These outputs resemble artistic works that can be protected by copyright. However, the AI itself is not a person who could claim copyright over the work, and it's not clear that there is anyone with the standing to claim to be the author of the work and claim such a copyright. At the moment, the legal landscape is unclear, and the Canadian government is currently consulting on issues surrounding copyright and AI. This issue is discussed in more detail in section 5.4 below.

3. AI Legislation and Regulation

3.1 Introduction

A number of jurisdictions have set out plans to create rules around AI technologies. There are a number of related reasons for this need to tailor some form of regulation to AI in particular. One is to protect people and communities from potential risks arising from AI systems. AI systems present some unique ethical challenges. For example, many such systems are so-called “black boxes”; with even the designers of some systems unable to fully explain how they arrive at their outputs. Related to this is the autonomy of many such systems, where they are capable of potentially making decisions that can have large impacts on individuals or groups. There have also been numerous examples of such systems incorporating biases against certain groups. Some of the harms of AI fall under existing laws and regulations, such as consumer protection laws, privacy laws, or sectoral specific regulations such as rules around medical devices. However, some of these harms slip through the cracks of such rules, and often the enforcement of these existing rules is difficult in the case of AI. Rules tailored specifically to AI can better protect against the ethical challenges raised by such systems.

Another reason to want specific rules around AI systems is to support innovation and investment in such technologies. The current legal and regulatory landscape can be quite unclear when it comes to AI, and this can deter innovation and investment. Having clear rules that avoid excessive regulatory burden can remove this uncertainty and incentivize beneficial developments in AI.

It is also valuable to ensure a consistency both within and across national boundaries. Within national boundaries, a single set of rules ensures that AI systems can be introduced across the country without worrying about complying with many potentially differing requirements. Internationally, having a set of rules that is broadly compatible with the rules being developed in other major markets ensures that AI systems developed in one country can be sold and deployed internationally.

As of now, only the European Union has passed specific legislation dealing with AI. However, legislation has been proposed in several countries, and other countries have developed non-legislative guidance that can either serve some of the same functions as legislation, and/or indicate the kinds of concerns that we can expect to see reflected in legislation when it is proposed. The most salient of these approaches for our purposes will be those of Canada, the United Kingdoms, the United States, and the European Union. Below we will look at the approaches in each of these jurisdictions, followed by a comparison between Canada’s proposed law and the approaches in the other jurisdictions.

3.2 The Artificial Intelligence and Data Act (AIDA)

AI technology in Canada is currently regulated by a number of different overlapping laws. The Personal Information Protection and Electronic Documents Act (PIPEDA) provides important guardrails around how businesses use personal information, while the Canadian Human Rights Act and the Criminal Code also apply to certain uses of AI. Within specific sectors, AI also falls under sectoral regulations such as the Bank Act, and Health Canada has issued guiding principles for the development of medical devices that use machine learning. However, there is not currently any laws that apply specifically to AI technologies; AIDA, if passed, would fill this gap. This is needed because not all uses of AI fall neatly into existing legal categories, and while human rights law, for example, would apply in cases of algorithmic discrimination, it can be difficult to become aware that one has been discriminated against in this way, and still more difficult to prove it.

The laws stated purpose is to establish common requirements throughout Canada for the design, development, and use of Artificial Intelligence in the private sector and to prohibit conduct in relation to AI systems that could result in serious harm to individuals or their interests³⁸. The plan is that introducing such a law will, on the one hand, provide a plan to manage the risks of this emerging technology and maintain public trust in Artificial Intelligence technologies, while also on the other hand, resolving regulatory uncertainty for AI researchers and innovators, who previously may have been unsure what regulations apply to their work.³⁹ Added clarity will help these researchers and innovators to continue their work with confidence about the regulatory landscape within which they are working.

On November 28th, 2023, the Minister of Innovation, Science and Industry François-Philippe Champagne presented a series of proposed amendments to AIDA to the Standing Committee on Industry and Technology, which seek to respond to feedback received from consultations.⁴⁰ Since these amendments are not yet part of the official bill, the proposed amendments will be covered alongside the current text of the bill.

AIDA defines an artificial intelligence system as “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.”⁴¹ The proposed amendments would revise this definition to “a technological system that, using a model, makes inferences in order to generate output,

³⁸ Bill C-27, 2021

³⁹ ISED, 2023

⁴⁰ Champagne 2023

⁴¹ Bill C-27, 2021

including predictions, recommendations or decisions.”⁴² This would align the definition more closely with the definition adopted by the Organization for Economic Cooperation and Development (OECD), and focuses less on particular technologies, which makes it more robust to potential new developments in AI.

However, the act itself does not apply to all AI systems. Instead, it is restricted to applying to “high impact” AI systems. The act itself does not define high-impact systems; it leaves that task to future regulations. Regulations provide support to laws and are enforceable by law but are not passed by parliament. Instead, regulations are crafted and put in place by persons or bodies that Parliament has given the authority to make them in an Act, such as the Governor in Council or a Minister. This makes regulations easier to create and amend than legislation. That being said, new regulations still require a period of consultation and stakeholder feedback, as well as drafting and review procedures, so they can still take some time to implement.

On the one hand, this makes the act more flexible, since it can be extended to include new categories of AI as they emerge or as they become potentially problematic without the need for further legislation. On the other hand, it leaves the law as it stands incomplete, and requires that the law be passed without knowing to what it applies. In addition, the drafting of such regulations is estimated to take an additional two years. With the speed at which AI technology is right now evolving, this delay makes the law look less agile and responsive than it at first appears.

The proposed amendments to the bill would address some of these worries. It would create a schedule of high-impact uses of AI that would be added to the bill. The Governor in Council would have the power to add to, modify, or delete any of these categories by enacting regulations. This would allow the categories of high-impact systems to be flexible over time, while meaning that there would not be a delay of up to two years to learn which AI systems the law applies to.

There are seven proposed categories of high-impact systems. The first is the use of AI in hiring, including determining remuneration, promotion, or termination. The second is the use of AI in determining whether to offer a service to someone, or the price of a service, or the prioritization of services offered. The third is the use of AI to process biometric information, including the identification of individuals and the assessment of an individual’s behaviour or state of mind. The fourth is the use of AI in content moderation or the prioritization of certain content. The fifth is the use of AI relating to health care or emergency services. The sixth is the use of an artificial intelligence system by a court or administrative body in making a decision about an individual. The

⁴² Champagne, 2023

seventh is the use of an artificial intelligence system to assist a peace officer in the exercise and performance of their law enforcement powers, duties, and functions.

This covers much of the problematic uses of AI. As Teresa Scassa notes, it conspicuously leaves out rental accommodations, which are generally not treated as a service in Canadian law, and the rules around biometric data seem unduly narrow. They only cover identifying individuals and assessing their state of mind via biometric data and leaves out the ability to identify individuals and assess an individual's state of mind via other means such as geo-location and IP addresses, purchasing habits, or browser history and cookies.⁴³

The bill contains a number of requirements on those who operate AI systems. Anyone who is responsible for an artificial intelligence system must assess whether it is a high-impact system. Those that do fall into the category of high-impact systems must then establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system, and establish measures to monitor compliance with these mitigation measures and their effectiveness. They must also keep records relating to both of these activities, and their manner of anonymizing data. Anyone developing or operating a high-impact system is subject to certain transparency requirements. They must publish on a publicly available website a plain-language description of the system that includes an explanation of how the system is used or is intended to be used; the types of content that it generates or is intended to generate and the decisions, recommendations or predictions that it makes or is intended to make; the mitigation measures established in respect of it; and any other information required by regulations. Finally, a person who is responsible for a high-impact system must notify the Minister of Industry if the use of the system results or is likely to result in material harm. What qualifies as material harm is a matter that is left up to future regulations to specify.

The bill provides for administrative monetary penalties for violations of the act. The act states that these are not intended to punish, but to promote compliance with the act. As yet undrafted regulations will establish the extent of these administrative monetary penalties. There is also the option for the minister to treat a breach of the act as an offense, and proceed via prosecution, where penalties can be up to \$10,000,000 or 3% of the company's gross global revenues in its financial year before the one in which the company was sentenced, whichever is greater.

The Minister of Industry is charged with oversight for the act. They have several powers to investigate AI systems for compliance with the law, including requiring the person to conduct either an internal or an external audit with respect to the possible contravention. The Minister may also order a person responsible for a high-impact system to cease

⁴³ Scassa, 2023a

using it or making it available for use if they have reasonable grounds to believe that its use gives rise to a serious risk of imminent harm.

One concern about AIDA is that the enforcement of the bill's provisions is centered in the same ministry that is responsible for drafting the law and the accompanying regulations. Thus, there is no independent regulator who would be called on to monitor and enforce the bill. This is a further cause for worry because the same ministry that is charged with enforcing AIDA is also charged with supporting the AI industry in Canada, creating a mixed set of incentives for rigorous enforcement.

The proposed amendments to the bill try to address these worries by expanding the role of the AI and Data Commissioner, a new role created by the bill. The amendments move many of the key oversight roles from the minister to the commissioner. However, the commissioner is an employee of the minister, chosen by them and able to be replaced by them at will. As such, this is still far from an independent oversight body, and the core conflict of interest identified above remains.⁴⁴

The key provisions of the bill are about mitigating biased output and harm. Biased output is defined in terms of unjustified differential treatment 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. The bill defines harm, meanwhile, in a distinctively individualistic fashion. It states that “harm means (a) physical or psychological harm to an individual; (b) damage to an individual’s property; or (c) economic loss to an individual.”⁴⁵ There is no provision, therefore, to recognize harm to a group or collective, unless such harm can be reduced to harm to individuals. This is a concern for First Nations in particular, since misuses of First Nations data and misrepresentation of First Nations communities can constitute distinctively collective harms that this law would not address. The law does prohibit “biased output” as well as harm, which may fill in some of these gaps, but does not address all of them. Biased output, after all, is also defined in terms of differential treatment of an individual and does not include concerns of group rights as such.

The proposed amendments would also add rules applying to “general purpose AI systems.” This is defined as “an artificial intelligence system that is designed for use, or that is designed to be adapted for use, in many fields and for many purposes and activities, including fields, purposes and activities not contemplated during the system’s development.”⁴⁶ The rules would apply to these systems whether or not they were in the category of high-impact systems, although a system could be both general purpose and high impact. These rules are partially intended to address the rise of generative AI

⁴⁴ Scassa, 2023b

⁴⁵ Bill C-27, 2021

⁴⁶ Champagne, 2023

systems, such as chatGPT (see section 2 for more details on generative AI). For example, they would require that if the system generates digital output consisting of text, images or audio or video content, then members of the public could identify the outputs as having been generated by artificial intelligence, either on their own or with the help of freely available software. In addition, there are other requirements on general-purpose systems, including carrying out an assessment of the adverse impacts that could result from any use of the system that is reasonably foreseeable, and measures to mitigate these adverse impacts having been established and the effectiveness of these measures assessed.

The AIDA makes no specific mention of First Nations or other Indigenous groups. The same is true of the wider privacy bill, The Digital Charter Implementation Act, of which AIDA is a component. This is important because AI makes extensive use of data, which can potentially include First Nations data of various kinds. As noted by Lisa Austin and John Borrows, nowhere in the law is there recognition of the special protections required for First Nations data, such as a recognition of OCAP® or other recognition of data sovereignty.⁴⁷ This is a huge hole in the regulatory aims of the law, ignoring the rights of First Nations over their data.

Additionally, the Assembly of First Nations has submitted a brief to the Parliamentary Standing Committee on Industry and Technology in October of 2023 that is sharply critical of the bill, both in the process by which it has been crafted and the substance of the bill.⁴⁸ The process based complaint is that the government did not engage in Nation-to-Nation consultation with First Nations during the drafting of the bill. While there was broad consultation with stakeholders, this falls short of the requirement to consult with First Nations on laws that will affect them. As the brief puts it “stakeholders are not rights holders, no matter how expert or experienced and limited engagement with some Indigenous individuals or representatives does not meet the legal and moral duty to conduct Nation-to-Nation consultation.”⁴⁹ Thus, the process used to draft the bill violates First Nations data sovereignty (discussed in more detail in section 5). It also breaches Canada’s obligations under the UNDRIP act, which commits Canada to upholding the principles of the UN declaration on the rights of Indigenous peoples.

On the substance of the act, the brief objects to the lack of independent enforcement in the bill, as was pointed out earlier. In addition, it objects to the exemptions contained within the act, which include National Defence, the Canadian Security Intelligence Service, and the Chief of the Communications Security Establishment, and “any other

⁴⁷ Burrows and Austin, 2022

⁴⁸ Brief by the Assembly of First Nations to the Parliamentary Standing Committee on Industry and Technology, 2023

⁴⁹ Brief by the Assembly of First Nations to the Parliamentary Standing Committee on Industry and Technology, 2023

person who is responsible for a federal or provincial department or agency and who is prescribed by regulation”.⁵⁰ The brief worries that other federal organizations that work with First Nations could be similarly exempt, such as Crown-Indigenous Relations and Northern Affairs Canada or the RCMP, and as such use potentially harmful AI systems in matters that concern First Nations. Finally, the brief also objects to the individualistic focus of the act, and the exclusion of any protection from community harms or infringements of community rights.

3.3 The UK’s regulatory approach to AI safety:

The UK’s approach to regulating AI is highly focused on promoting innovation and investment in AI technologies within the UK. This is highlighted in the name of the white paper that lays out the plan, “A pro-innovation approach to AI regulation”, and repeated multiple times throughout the plan itself.⁵¹ Even when attention is given to fostering public trust in AI, this is grounded in concerns over innovation, stating that “If people do not trust AI, they will be reluctant to use it. Such reluctance can reduce demand for AI products and hinder innovation.”⁵² This concern for trusted AI is subtly but importantly distinct from a concern for trustworthy AI—given the stated goal of increasing innovation, this can be achieved as long as the public in fact trusts AI technologies, regardless of whether this is because these technologies are genuinely trustworthy or instead because the public has been fooled into thinking them trustworthy.

The UK’s approach, like AIDA, focuses on high-risk systems. As the white paper states, “we will ask that regulators focus on high-risk concerns rather than hypothetical or low risks associated with AI. We want to encourage innovation and avoid placing unnecessary barriers in its way.”⁵³

AI is defined in terms of two characteristics: its adaptiveness and its autonomy. Adaptiveness refers to the way that AI operates on the basis of instructions which have not been expressly programmed with human intent, having instead been ‘learnt’ on the basis of a variety of techniques. Autonomy is the extent to which these systems are able to operate without human oversight or intervention. It is not entirely clear from the description whether both of these characteristics need to be in place for every system that is classified as AI, or whether a high degree of one characteristic is sufficient. For example, consider a highly adaptive system, which uses machine learning to come to conclusions that the designers did not anticipate, nor can they explain. Would it be counted as AI if it had no autonomy, and only offered a prediction to a human who had the final say in the decision? For example, we could look at the use of AI to assign risk

⁵⁰ Bill C-27, 2021

⁵¹ Department of Science, Innovation and Technology, 2023

⁵² Department of Science, Innovation and Technology, 2023

⁵³ Department of Science, Innovation and Technology, 2023

scores for criminals. These systems attempt to predict which criminals will reoffend and are used by judges to inform decision about bail amounts, parole, and more. However, the AI does not make any decisions on how to treat defendants. It only offers a risk score, which the judge then can take into account in their own decision making. Thus, while the system is highly adaptive, it exercises little to no autonomy of its own. Or conversely, we can imagine a system with next to no adaptiveness, such as a simple spreadsheet into which values are added and a result produced, but that has a high degree of autonomy, so that the conclusions of the spreadsheet are implemented in high-stakes scenarios without human oversight. It is unclear if this would count as AI on the UK's definition.

Instead of seeking to pass laws regulating AI, the UK intends to empower existing regulators to regulate AI that comes within their purview, while being on the lookout for uses of AI that may slip through the cracks so that existing regulators can be empowered to regulate these uses or legislative action can be taken. The government's role, then, is to play a coordinating function between these different regulators to ensure both a harmony of approach and that no high-risk uses of AI fall between existing regulators jurisdictions. However, they stress that the context specific nature of this approach, with potentially different rules for different sectors, is appropriate and a strength of the approach; since "AI is a dynamic, general-purpose technology and ...the risks arising from it depend principally on the context of its application."⁵⁴

The plan proposes six cross-sectoral principles for the regulation of AI. These are:

- 1) Ensure that AI is used safely
- 2) Ensure that AI is technically secure, and functions as designed
- 3) Make sure that AI is appropriately transparent and explainable
- 4) Embed considerations of fairness into AI
- 5) Define legal persons' responsibility for AI governance
- 6) Clarify routes to redress or contestability

3.4 United States guidance on AI

The United States as of now has no proposed laws around AI regulation. However, there have been some non-binding documents produced that give insight into the government's current thinking on AI regulation, and as such give us a hint of how they may proceed. The most substantial of these are the National Institute of Standards and

⁵⁴ Department of Science, Innovation and Technology, 2023

Technology (NIST) Artificial Intelligence Risk Management Framework⁵⁵ and the White House Office of Science and Technology Policy's Blueprint for an AI Bill of Rights.⁵⁶ In addition, on October 30th, 2023, the White House issued an executive order on safe, secure and trustworthy artificial intelligence.⁵⁷

The NIST framework defines AI as “an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy”.⁵⁸ It adapts this definition from the Organization for Economic Co-operation and Development (OECD) recommendations on AI.

The NIST framework identifies the characteristics of trustworthy AI systems as including valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. However, simply identifying these criteria is not enough; each AI system will need to balance these different demands based on the context in which it is put into use.

The NIST framework identifies four key functions of a risk management framework: Map, Measure, Manage, and Govern. Govern refers to the idea that a culture of risk management is present within the organizations. This is then realized through the other three requirements: Mapping risks related to the particular context in which an AI system is being deployed, measuring identified risks in order to track and assess these risks, and managing these risks by prioritizing and acting to mitigate these risks.

Along with the NIST framework, the White House Office of Science and Technology Policy has also issued its own document, the Blueprint for an AI Bill of Rights. This document proposes the following set of principles for an AI bill of rights:

- 1) You should be protected from unsafe or ineffective systems
- 2) You should not face discrimination by algorithms and systems should be used and designed in an equitable way.
- 3) You should be protected from abusive data practices via built-in protections, and you should have agency over how data about you is used.
- 4) You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.
- 5) You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter

⁵⁵ National Institute of Standards and Technology 2023

⁵⁶ White House Office of Science and Technology Policy 2022

⁵⁷ Biden 2023

⁵⁸ National Institute of Standards and Technology 2023

This framework is meant to apply to automated systems that have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services. It defines automated systems as "any system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities. Automated systems include, but are not limited to, systems derived from machine learning, statistics, or other data processing or artificial intelligence techniques, and exclude passive computing infrastructure."⁵⁹

The Blueprint for an AI Bill of Rights makes special mention of the need to protect communities as well as individuals. The authors recognize that U.S. law has been better at protecting individual rights than it has been at recognizing effects that emerge most clearly at the community level and acknowledges the need to explicitly recognize and protect against these kinds of harms.

The executive order directs a range of actions to be taken to improve the oversight on AI systems being developed and used. The executive order contains a variety of directives to different organizations. One component requires companies developing any foundation model that poses a serious risk to national security, national economic security, or national public health and safety to notify the federal government when training the model and share the results of safety tests. This applies only to a small subset of AI systems but does create a transparency requirement for these models. Similarly, the National Institute of Standards and Technology will set the rigorous standards for extensive testing to ensure safety of AI systems before public release. However, this will not apply to all AI systems. Instead, the Department of Homeland Security will apply those standards to critical infrastructure sectors and establish the AI Safety and Security Board. The Departments of Energy and Homeland Security will also address AI systems' threats to critical infrastructure, as well as chemical, biological, radiological, nuclear, and cybersecurity risks. Similarly, the Department of Health and Human Services will establish a safety program to receive reports and act to remedy harms or unsafe healthcare practices involving AI. These are important first steps in regulating AI but restricted to specific areas of special importance.

In the area of privacy, the government commits to funding privacy preserving technologies and assessing the government's own use of information but doesn't put any restrictions on the private sector. For equity and civil rights, the executive order directs that clear guidance be provided to landlords, federal benefits programs, and federal contractors to keep AI algorithms from being used to exacerbate discrimination. The directive attempts to address broader issues of algorithmic discrimination through

⁵⁹ National Institute of Standards and Technology 2023

training, technical assistance, and coordination between the Department of Justice and Federal civil rights offices on best practices for investigating and prosecuting civil rights violations related to AI. Finally, the executive order tries to ensure fairness throughout the criminal justice system by developing best practices on the use of AI in sentencing, parole and probation, pretrial release and detention, risk assessments, surveillance, crime forecasting and predictive policing, and forensic analysis. Again, we see an important recognition of the issues, but restricted to specific sectors and applications.

The executive order also contains components relating to supporting workers who might be displaced or otherwise harmed by AI, promote innovation and competition, advance American leadership abroad, and ensure responsible and effective government use of AI. The executive order has multiple components and acts to mitigate a range of risks of AI. However, likely because of the limitations of executive orders as opposed to legislation, the approach is somewhat fragmented and largely does not address the private sector outside of particular high-risk areas where government has special regulatory authority.

3.5 European Union Artificial Intelligence Act

The European Union is the first jurisdiction to pass a law specifically addressing AI risks. The Artificial Intelligence Act was approved by the European Parliament on March 13th, 2024, by a vote of 523 votes in favour, 46 against and 49 abstentions.⁶⁰ The law is risk based, dividing AI systems into four categories: unacceptable risk, high risk, limited risk, and minimal risk. Unacceptable risk systems will be banned outright, with high-risk systems having the most stringent regulations, and other systems having lighter restrictions mostly related to transparency.

The EU act provides the following definition of AI; “Artificial intelligence system’ (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate output such as predictions, recommendations, or decisions influencing physical or virtual environments.”

The unacceptably high risk systems, which would be forbidden under the proposed law, include systems that use subliminal techniques in a way that can cause physical or psychological harm, “social scoring” style systems that would rate individuals on their social behavior and then apply this to their detriment in other contexts, and the use of real-time remote biometric identification systems for law-enforcement outside of some carefully circumscribed use cases.

⁶⁰ 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, 2021

High risk systems are those that are either (a) AI systems required to undergo an EU conformity assessment process, or systems that are safety components of a product required to undergo such an assessment; or (b) AI systems that fall into one of the specified high-risk use cases. The specified use cases are contained in an Annex to the law that can be added to without the need to change the law itself, making this definition more flexible. The current high-risk use cases include all uses of biometric identification, AI used as safety components in critical infrastructure such as the management and operation of road traffic and the supply of water, gas, heating and electricity, AI systems used in recruitment, termination, or promotion decisions, AI systems used to detect the emotional state of individuals in areas such as law enforcement or immigration, and AI used in the determination of eligibility for benefits.

High-risk systems have a number of requirements placed on them. They must have a risk management system in place that aims to identify and mitigate the potential risks of such a system. They must also be tested to ensure they follow the requirements and to identify appropriate risk management approaches. There are also requirements placed on the data used to train, validate, and test AI. These data must meet a number of requirements, including that they be examined in light of possible biases, and the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed. There are also rules around technical documentation and record keeping for high-risk AI systems. There is also a requirement for human oversight over high-risk systems, whereby a human being is positioned to be able to intervene and countermand the decision made by a high-risk AI system. Finally, high-risk AI systems must have an acceptable degree of accuracy, robustness, and cybersecurity.

Systems that are not high-risk do not need to follow all of the above requirements. They do need to follow transparency requirements, whereby individuals are informed when they are dealing with an AI system, and when content has been altered by an AI system (as in “deep fakes”). They are also encouraged to voluntarily sign on to codes of conduct that contain some or all of the requirements applied to high-risk systems, and potentially other requirements having to do with things like environmental sustainability.

Penalties focus on administrative fines, with different levels of fines for violations of different components of the act. The steepest of these administrative fines is up to €40 million, if the offender is company, up to 7 % of its total worldwide annual turnover for the preceding financial year, whichever is higher. This penalty is for use of types of AI systems prohibited as being of unacceptable level of risk, or for violation of the data and data governance provisions of the act. Other substantive provision will be subject to fines of up to 4% of worldwide turnover or €20 million, with a lower category of fines of up to 2% of worldwide turnover or €10 million for administrative failures in dealing with enforcement bodies.

The enforcement mechanism for the act is similar to that for the General Data Protection Regulation (GDPR). Each member state will be required to designate or create a national regulatory body that will enforce the AI act. The European Commission, meanwhile, will co-ordinate issues that effect the European Union as a whole, advised by a new AI council. There is currently some debate about whether more supervision and enforcement should be carried out centrally by the Commission and a newly formed European Artificial Intelligence Office.

3.6 Comparing AIDA to international approaches

The proposed Canadian law governing AI, as we have seen, exists alongside a number of alternative approaches. So how does it compare to these other approaches? We will start by comparing the Canadian approach to the U.S., then the UK, and finally the European Union AI law.

The United States has yet to propose an actual piece of legislation to regulate AI. Instead, they have relied on an executive order, along with the risk management framework and blueprint for an AI bill of rights. These are more limited than the Canadian AIDA, being focused on non-binding guidance and rules directed to specific sectors. This is the limitation of the use of executive orders rather than legislation, since the authority is limited to what the president has jurisdiction over.

However, despite the overall more limited approach to AI regulation that the US has adopted, it is notable that they have explicitly acknowledged the need to address community level harms. This stands in stark contrast to the AIDA, which explicitly defines harm in individualistic terms and overall focuses on the effects of AI on particular people, leaving out community or societal level harms. This is an area where the Canadian approach could stand to learn from the US.

Both the Canadian and the UK approach begin with the same two goals: supporting innovation and building trust in AI. Both also attempt to be “agile” approaches, to be able to adapt to the rapid pace of change in the technology underpinning AI and the likelihood that new issues will arise over time. However, the way they go about this goal is very different. The AIDA does this by creating a legislative structure, while leaving many important details about how the legislation will function up to the regulations. The idea is that regulations are more flexible than the legislative components of the law and can thus be changed to more appropriately adapt to change. The UK, on the other hand, looks to avoid legislation altogether. Instead, they rely on existing regulators, seeking to encourage and empower them to regulate AI. They will also encourage these regulators to point out areas where there are gaps in the regulator’s authority, and to propose ways to address them.

The UK’s approach has the advantage of being more flexible than the Canadian legislation. While regulations are more flexible than legislation, they are still time

consuming and difficult to adopt and change. It is estimated that it will take up to two years to draft the regulations needed for the AIDA to come into force. On the other hand, guidance to regulators can be adapted and changed much more quickly, and regulators can be given the ability to use their discretion, giving more flexibility. However, the downside is that there may end up being large gaps in the authority of regulators, which may take time and effort to address. The AIDA is able to apply to a broader range of situations, without the need to coordinate and fill in gaps in the authority of existing regulators.

In any case, the approach taken by the UK is likely not available to Canada, due to Canada's division of powers between the provinces and the federal government. Some of the regulators, and some of the regulatory gaps, are under provincial jurisdiction. So, to empower regulators or to address these gaps would require coordination with each of the provinces, rather than being a top-down exercise of federal control. Thus, the legislative approach with details left to regulation may be best option for an agile approach to AI regulation for the Canadian context.⁶¹

Of all the approaches looked at in this chapter, the AIDA most closely resembles the EU Artificial Intelligence act. Both take a legislative approach to regulating AI, and they use similar mechanisms to enforce their rules, such as administrative monetary penalties. However, the EU law applies not just to high-impact systems, but to all AI systems. While high-risk systems are singled out for additional restrictions and requirements, other forms of AI systems are equally covered by the law. In addition, the EU law leaves less to later regulations than AIDA does. As we saw earlier, the AIDA does not even tell us what AI systems it applies to, since the definition of high impact systems is to be determined by regulations, although this is addressed in the proposed amendments. In addition, the EU law is enforced by an independent body, with member states creating a national regulatory body to enforce the law. This is in contrast to AIDA, which vests enforcement powers in the Minister of Industry, or an AI and data commissioner employed by and reporting to the Minister. This is the same Minister responsible for both the law itself, and for growing the Canadian AI industry.

On the other hand, the EU approach is likely to be less flexible than the Canadian approach. Since more of the law is fixed by legislation, rather than regulations, it will require legislative change to adapt these elements of the law. This can be especially important given rapid pace of change within the AI industry, and the possibility that new technologies or applications will change what we might want to classify as high-risk.

⁶¹ Scassa 2023c

4. Discrimination by AI

4.1 Introduction

One pressing concern about the ever-wider adoption of AI is discrimination by AI systems. This refers to instances where an artificial intelligence system produces outputs that are biased against certain groups, especially historically marginalized groups. This is often referred to as “algorithmic bias”, or “algorithmic discrimination”. First Nations have long been subject to discriminatory and unfair treatment, and the potential for this treatment to continue or even be exacerbated by AI systems is of great concern.

There are numerous documented instances of algorithmic bias in the real world. One of the most famous is the COMPAS system used by the U.S. penal system to provide “risk assessments” of inmates, assessing how likely they are to reoffend. This was used in making decisions about bond amounts, parole, and even sentencing. Propublica, nonprofit organization dedicated to investigative journalism, found that the system was much more likely to provide false positives for Black people, labelling them high risk when they did not end up going on to reoffend, and false negatives for white people, labelling them low risk when they did in fact end up reoffending.⁶²

Another example is the case of an AI system used in the U.S. to ascertain medical need in patients, and thus to assign additional care to those who are sicker. This system used medical spending as a stand in for medical need. However, within the U.S. health system less money was spent on black patients than on white patients with equivalent medical need. Therefore, the AI system trained on this data systematically assigned lower medical need scores to black patients than to white patients with equivalent needs, which prevented black patients from accessing the additional health care resources they needed.⁶³

An initially promising approach to mitigating algorithmic bias is to ensure that sensitive attributes are not used as part of the input to the AI system. For example, a company might ensure that no race data is used as an input to a system deciding on loan applications. However, an AI system might be able to infer these omitted characteristics from the information it is provided. For example, postal codes can be highly correlated with race, and might be used by an AI system as a proxy for race. A real-world example of this comes from Facebook’s ability to target ads to members of specific races, even though race data is not collected by Facebook. Facebook permitted advertisers to target ads for housing, for example, to certain “ethnic affinities”, which were determined by looking at pages and posts the users have liked or engaged with.⁶⁴ In this case, the

⁶² Angwin et al., 2016

⁶³ Obermeyer et al. 2019

⁶⁴ Angwin and Parris, 2016

decision to target advertisements by race was intentional on Facebook's part, but an AI system could end up using proxy variables to make decisions based on discriminatory characteristics without any such intention, and indeed without the programmers even being aware that this was happening. See the discussion of black-box AI in section 4.3.1 for more about how this can occur. Therefore, simply removing explicit data about race or other characteristics is insufficient to prevent algorithmic bias. As AI systems are adopted for an ever-wider array of tasks, the risks and consequences of algorithmic bias similarly expand.

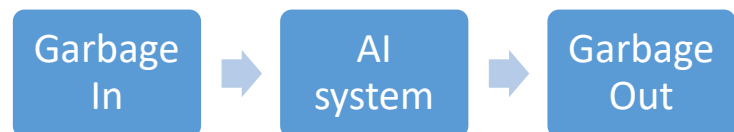
4.2 Causes of Algorithmic Bias

Algorithmic bias can be the result of a number of different factors. Some of these depend on the design of the algorithm itself, while others are due to the training data used. Below we will look at some common causes of algorithmic bias.

4.2.1 Bias from training data

Biases in training data will always pose a risk for perpetuating algorithmic bias. A machine learning system is only as good as the data that it has been trained on. As the common data science saying goes, garbage in means garbage out. If the inputs are flawed, then no algorithm will produce reliable outputs.

One way that training data can lead to algorithmic bias is when a marginalized group is not included in the training data or makes up too small a proportion of the data. This can result in worse results for the marginalized group, since the AI system will not have learned how to assess members of this group. As an example, facial recognition systems tend to be less accurate for women, for Black people, and especially for Black women.⁶⁵ This is probably due to the fact that the training data used for these AI contained more instances of white and male faces, and so the systems learned to classify such faces most accurately, while an insufficient representation of black women lead to the system making many more errors in such cases.



A second way that training data can contribute to algorithmic bias is if the training data is itself partially the result of human bias, which can then be taken up and perpetuated by the AI systems trained on such data. As an example of this kind of algorithmic bias, we can look at instances of google search autocomplete. Several discriminatory

⁶⁵ Buolamwini and Gebru, 2018

autocomplete suggestions have been discovered and pointed out, and while google often fixes the specific searches that are called out in articles, others can still be found. Examples include autocompleting “are Jews” to “are Jews evil?” (with most of the search results claiming that the answer is yes), and “blacks are” to “blacks are not oppressed” and “feminists are” to “feminists are sexist”. The google autocomplete AI system learns from past searches, and so these biased results are caused by examples of bias in the data being used to train the system.⁶⁶

Another example, where the bias is more easily missed, is provided by the AI amazon developed for its hiring process. This AI produced a rating when provided with a resume of a prospective hire. It was trained on the resumes submitted to the company over a 10-year period. However, this data itself encoded the sexism of society, since more men than women were hired by Amazon, especially in tech roles. This resulted in an AI hiring system that was biased against women, downgrading resumes that contained the word “women’s” (as in women’s chess team) and graduates from two all-women’s colleges.⁶⁷

More subtly, the bias may enter in how the training data is labelled. To learn from data, most AI systems need the data to be labelled; for example, in order to learn how to recognize birds an AI system would need to have a large number of pictures labelled as to whether or not they contain birds. These labels are assigned by human beings, and thus the act of labelling can result in algorithmic bias. For example, we can imagine an AI system that is being trained on labelled images of people, to identify various traits. However, if the labelers are more likely to associate positive traits with, say, white males, and negative traits with women and marginalized groups, then the resulting algorithm will learn this same biased pattern of evaluation.⁶⁸

4.2.2 Bias from algorithm design

Along with bias that is caused by the training data used to develop AI systems, the design of the algorithm itself can also be a source of bias. This could be the result of errors in programming, but more commonly it will arise as a result of the choice of what data to use and how to weight that data.⁶⁹

The choice of what data to sample, and how to prioritize it, is crucial in the design of artificial intelligence. This is especially true when it is difficult or impossible to directly measure the desired goal of the AI model. For example, we might want our AI to identify which patients are the most ill, and thus in the most need of additional medical resources. However, we cannot directly measure illness. Instead, we will need to

⁶⁶ Cadwalladr, 2016

⁶⁷ Vincent, 2018

⁶⁸ Dai and Brown, 2020

⁶⁹ To assign weight to data is to decide how much that piece of data will count towards deciding the result. The higher the weight, the more that piece of data will be taken into account in arriving at an outcome.

identify a proxy variable that we can measure. As we saw in the example provided earlier, this choice can easily end up encoding biases. In the case discussed by Obermeyer et al. (2019), health care spending was used as a proxy for degree of medical need. Since Black patients had historically had less money spent on them than white patients of an equivalent level of medical need, this led the algorithm to assign Black patients a lower risk score, meaning they received less health care. Similarly, an AI designed to evaluate teaching outcomes might use standardized test scores as a proxy for academic achievement. However, this choice would penalize schools with a greater proportion of high-needs students.

4.3 Risk Factors for Algorithmic Bias

4.3.1 Black-Box AI

As was discussed in section 2, many machine learning systems are so-called black boxes. This refers to AI systems in which the designers of the system cannot explain the outputs of the system. The output can be measured for accuracy, but the process by which the AI takes in data and arrives at conclusions cannot be satisfactorily explained. This can have a number of consequences, but one of them is to increase the difficulty of identifying algorithmic bias.

For a symbolic AI system, such as the expert systems discussed in section 2 above, discrimination can be demonstrated by pointing to the discriminatory programming of the system. For example, if an expert system is programmed to downgrade someone's loan application based on their postal code, and we know that postal codes are strongly correlated with race, then we can point to this as a cause of bias. However, this is much more difficult with black box systems. No one fully understands what factors the AI is taking into account, and so it is more difficult to point to causes of bias. If a series of individuals from historically marginalized groups are each denied a loan, it is difficult for them to know if this is a coincidence or a case of bias.



Black Box AI=
Humans Cannot
Explain Outputs

Algorithmic bias for black box systems must usually be based more on a pattern of results, rather than pointing to specific discriminatory features of the programming. This is how bias was identified in many of the machine learning examples discussed earlier. However, this can require a large body of examples to analyse, and thus the extended use of a biased system before its bias can be identified.

Many AI systems are proprietary, exacerbating the issues with black-box AI systems. If outsiders are not permitted access to the code of an AI system, it becomes more difficult to test hypotheses about how and whether a system is biased. Requiring third-party audits, where these auditors have access to the code of the system, can help to mitigate this, as discussed below in section 4.4.1.

4.3.2 AI Designed Without Diverse Input

Another major risk factor for biased AI is when the design and implementation of the AI system does not include diverse voices. If systems are designed by a homogenous group, they are much more likely to miss potential sources of bias. The technology sector can be worryingly homogenous at times, with a much larger proportion of white males than the general population. This increases the risk that AI systems will be designed by teams that fail to adequately represent or consider the effects of these systems on historically marginalized groups.

An excellent example of this is the discovery by Joy Buolamwini of the bias in commercial facial recognition software. She discovered this first because AI facial recognition software failed to register her face, which led to her discovery of its inaccuracy for black women in particular.⁷⁰ A more diverse development team would be more likely to discover this kind of flaw just from their own interactions with the software and be more likely to be aware of and sensitive to the dangers of this kind of bias.

A diverse development team, or at least one that consults with diverse stakeholders, will be more able to detect and work to prevent bias from all of the sources discussed earlier. People with diverse perspectives will be more likely to be aware of potential biases in the training data being used. They will also be more likely to identify problems with the data being selected or the proxy variables chosen for training AI systems. In addition, the lived experience of such diverse voices is crucial for identifying the potential ethical issues that could arise from a particular AI system, helping to shape AI design to take better account of these issues.

The diversity of those developing AI has begun to improve. As an example, in 2011, 71.9% of new resident Computer Science bachelor's graduates in America were white. In 2021, that number dropped to 46.7%. However, it was still the case that 78.7% of new AI PhDs were male in 2021, only a small decrease from the 2011 rate. Many groups remain greatly underrepresented, with Hispanics making up only 5.1% and Black or African Americans 4%. Native American or Alaska Native students made up only .64%, a tiny fraction of the total graduates and well below their 2.09% share of the general population.⁷¹

There is growing appreciation of the need for increased diversity, both within academia and industry. Google, for example, puts out an annual diversity report that tracks their

⁷⁰ Buolamwini, 2023

⁷¹ Maslej et al., 2023. All of this data is from the US, I was unable to find comparable data on the diversity of Canadian computer science graduates.

efforts to increase the representation of minorities. In their 2023 report, they report that in 2022, they met their Racial Equity Commitment of increasing leadership representation of Black+, Latinx+, and Native American+ Googlers by 30%.⁷² Still, this is starting from a very low baseline—in 2018 only 2.5% of their workforce was black. Furthermore, the overall percentage of Native Americans employed by Google has shrunk from 1% to .8% since 2014.

4.3.3 AI Without a Human in the Loop

Another risk for the use of AI is situations in which there is no human in the loop. In other words, an AI system is permitted to make the final decision on some matter of importance, with no human available to approve the decision or be appealed to in case of errors. This increases the risk of algorithmic bias, by removing accountability for the decisions made by the AI systems. This risk is magnified if the AI system is used in secret, so that those subject to it are not even aware that AI has made the decision. If human beings are not in the loop on decision making, then the risks that algorithmic bias will fail to be detected increase.

Even if a human is available to be appealed to in case of errors, effort must be made to avoid undue deference to AI decision making. AI decisions can appear much more reliable than they actually are, due to the hype around AI and the overall impression that technology will be more reliable than human beings. Ironically, in some cases having AI that is explainable, in other words that is not a black box, makes this problem worse. A study from 2021 showed that people were more likely to blindly trust AI predictions when the AI system was explainable than when it was a black box.⁷³ Thus, for all of their other dangers, black-box systems may in some cases ensure a healthy skepticism of AI decision making.

4.4 Mitigating the Risk of Algorithmic Bias

4.4.1 Auditing of AI

Any AI system that has a risk of algorithmic bias should be audited to ensure that any bias is detected and eliminated. These audits should be carried out by a third-party firm, to ensure impartiality and to catch errors that may have been missed by the design team. Auditors should have full access to the AI systems code and be able to inspect both the inputs and the outputs of the system for potential sources of bias.

Audits are required not just at the creation of an AI system, but continually throughout its life cycle. Some AI systems continue to learn from data they acquire during use, meaning that bias can creep into the training data after the system has been in use for some time. Even systems that do not continue to update based on new information can have subtle biases that do not become apparent until the system has been in place for some time. In addition, the way an AI system is used in practice might create bias that was not visible at the time of creation, but only emerges in the context of use. For

⁷² Google Diversity Annual Report, 2023.

⁷³ Poursabzi-Sangdeh, Forough, et al., 2021

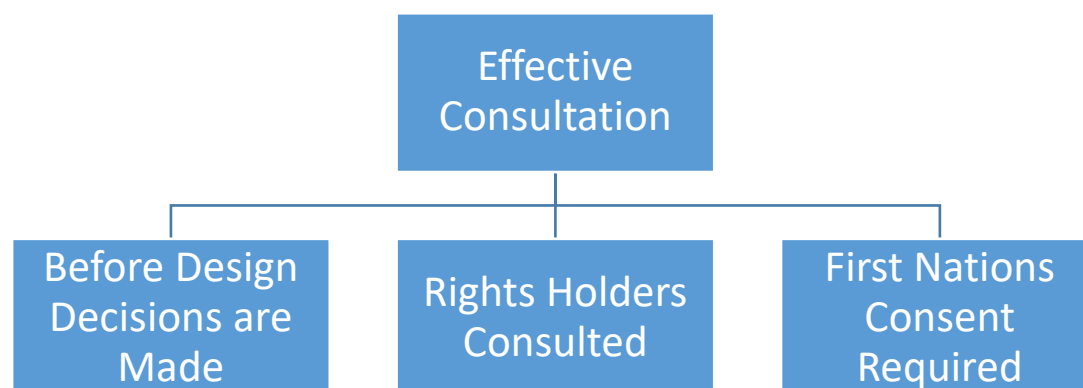
example, if doctors are more likely to blindly trust an AI in the case of racial minorities, but more likely to overrule the AI system and provide additional care in the case of white patients, then the AI system would be contributing to a biased system even if its outputs are identical between the two groups. Thus, auditing must be an ongoing effort, not a one-off exercise.

4.4.2 Consultation with First Nations

It is crucial that any AI system that might affect First Nations be developed in consultation with First Nations. As well as being required by data sovereignty considerations, discussed in section 5, this will also help in the detection and prevention of algorithmic biases.

It is important that this consultation begin from the outset of the project, when the overall aim and design of the AI system is being decided on, rather than being just a box to tick at the end of the design process. First Nations can provide crucial insight on the context of the data being used to train the AI model, which can identify bias that might otherwise be missed. In addition, First Nations are the best placed to identify when a proxy that is being used for the true goal of the AI system is likely to be inaccurate in ways that discriminate against First Nations. Finally, First Nations should have the ability to say when the use of AI is inappropriate, due to the risk of bias. All of these forms of feedback will be most useful at the beginning of a project, rather than once crucial decisions on shape of the AI system have already been made.

Consultation should take place with the rights holders—the First Nations themselves. Consulting with advisory tables or community groups is no substitute for consultation with the leadership of the affected First Nations. Ideally, this would also be accompanied by First Nations representation on the design team itself, in the form of First Nations individuals whose lived experience could help guide the design throughout.



4.4.3 Right to Object to AI Decisions

It is crucial that First Nations individuals and First Nations communities have avenues to object to AI decision making that could be the result of algorithmic bias. There are a number of ways that such a right to object could be realized. First and most crucially, it must always be transparent when and where AI systems are being used. There must be no AI used in secret. While this is a commitment of the current draft framework being proposed for the Ontario public services use of AI, this will also need to apply to private sector uses of AI. The federal AI and data act would, if passed, make it a requirement that those who make use of a high-impact system publish on a publicly available website a plain-language description of the system. However, there is no such provision for non-high impact systems, and the definition of high impact systems is left for regulations, leaving it unclear what AI might still be used in secret.

Alongside knowledge that AI is being used, an effective right to object must also allow for a right to have the decision made by AI reviewed by a human being. This is necessary if biased decisions are to be effectively challenged by those to whom they apply. However, an individual right to object is itself insufficient. It is often difficult to determine in individual cases whether algorithmic bias has influenced the result, and even harder to prove that this is the case. Many of the harms of algorithmic bias only become clear at the collective level. Furthermore, some harms may be collective in nature, and only legible when we look at communities. Thus, a right to collectively object to AI decision making is also required to head off the risk of algorithmic bias.

5. AI and Data Sovereignty

5.1 Introduction

As was discussed in the first chapter, machine learning makes use of huge quantities of data to train AI systems. There have been significant concerns about the rights to the data that is used. For example, a company called Clearview AI created a vast facial recognition database, which was used by police departments, by scraping 30 billion images from Facebook and other social media sites.⁷⁴ This raises serious privacy concerns. Even though the images used were publicly available, scraping the images to train AI, especially with the result being used by law enforcement to identify individuals as potential criminals, violated the social license for the use of these images and resulted in a great deal of anger towards both Clearview AI and Facebook for the way that personal information was treated. Similarly, generative AI systems such as ChatGPT have been trained using copyrighted materials, such as books and articles, while image generating systems such as DALL-E were trained using copyrighted images. Authors and artists are currently arguing in court that these uses of their intellectual property violate copyright, and that they have a right to be compensated for this use.⁷⁵

When we turn to First Nations data, these concerns about AI's use of data are magnified. First Nations data belongs to First Nations themselves, and its use in AI training risks violating the data sovereignty rights of First Nations. This includes personally identifiable data on First Nations individuals, but also community level data and cultural information, such as images of First Nations art and sacred objects or text recording First Nations traditional stories or teachings. In general, First Nation data sovereignty applies to "any facts, knowledge, or information about the nation and its citizens, lands, resources, programs, and communities."⁷⁶ This includes both data that is about First Nations, such as demographic, socio-economic, and health, housing infrastructure, and other services, as well as data from First Nations, such as traditional knowledge and languages.

Alongside the worries about data used to train AI, data sovereignty issues are also raised by the ability of AI to identify groups in data, even when these groups are not explicitly labelled. This raises the possibility that AI could be used to identify First Nations data within datasets even when those datasets do not contain explicit First

⁷⁴ Tangalakis-Lippert, 2023

⁷⁵ Appel at al., 2023

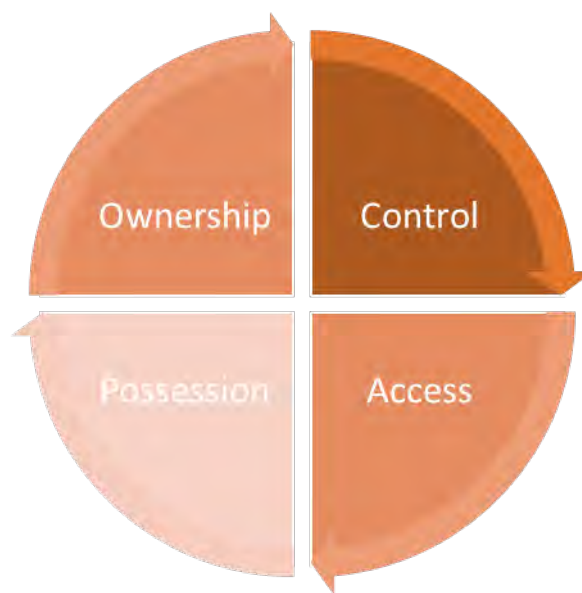
⁷⁶ Rainie, Rodriguez-Lonebear & Martinez, 2017

Nations identifiers. This potentially expands the range of datasets that raise data sovereignty concerns and creates worries over how this AI technology will be deployed.

5.2 First Nation data sovereignty

First Nations data sovereignty refers to the need for each individual First Nation to exercise control over the collection, management, and use of data concerning their community, peoples, land, and resources, in accordance with their worldview. First Nations data sovereignty is a fundamental right of First Nations and is also supported by several commitments made by the Government of Canada. It is supported by the United Nations Declaration of the Rights of Indigenous Peoples (UNDRIP), which United Nations Declaration on the Rights of Indigenous Peoples Act commits the Canadian government to implementing. Particularly relevant are articles 18 that states that Indigenous peoples have the right to participate in decision-making in matters which would affect their rights in accordance with their own procedures, and article 19 which holds that states are required to “consult and cooperate in good faith with Indigenous peoples through their own representative institutions in order to obtain their free, prior and informed consent before adopting and implementing legislative or administrative measures that may affect them”.⁷⁷ Applying these two sections to data, decisions about which clearly affect First Nations, and we see that this supports First Nations data sovereignty.

An important component of First Nations data sovereignty are the principles of OCAP®. The acronym stands for “Ownership, Control, Access, and Possession”. Ownership refers to the relation between First Nations rights holders and their data; it asserts that First Nations rights holders’ own data about them in the same way that individuals own data about themselves. Control states that First Nation rights holders have the right to control all aspects of the research and information management processes that involve them. This includes the collection, use, disclosure, and destruction of First Nations data. Access holds that First Nations must have access to their information, wherever it is held. It also asserts the right of First Nations rights holders to decide who else should



⁷⁷ United Nations, 2008

have access to the data. Finally, possession is the mechanism whereby ownership can be asserted and protected. This refers to the physical control over the data.⁷⁸

While we can outline what each of these principles mean, OCAP® itself outstrips the definition of each individual word in the acronym. It represents First Nations principles and values of data sovereignty, and the four elements are part of an inseparable whole, where no one of the elements can exist without the others. As Bonnie Healey put it, “We cannot ignore ‘ownership’ or ‘possession’ any more than the Four Directions can omit the East or the North.”

The national First Nations Data Governance Strategy, developed by the First Nations Information Governance Centre and endorsed by the Chiefs in Assembly, clearly lays out that rights holders are the appropriate authorities for exercising First Nation control over data.⁷⁹ Rights holders are the First Nations governments, and through them First Nations citizens themselves. Thus, for the collection or use of data about a specific community, a band council resolution would be an appropriate way to ensure First Nation control over their data, while for province wide data use, a resolution from the Chiefs in Assembly would be an appropriate form of control. As we shall see in the next section, data sovereignty in general and OCAP® in particular are threatened by the way that many machine learning systems are trained.

5.3 AI training data and data sovereignty

As we saw in chapter 2, machine learning AI systems require huge quantities of data for their training. This has led to the rise of what is sometimes referred to as “big data”. Big data is often discussed in terms of data that possesses the five V’s: velocity, volume, value, variety, and veracity.⁸⁰ Velocity is the speed at which the data is created and how quickly it can be accessed. Volume refers to the sheer amount of data available. Value refers to the data being meaningful and useful. Variety refers to the data being diverse and collected from a wide range of cases. Finally, Veracity refers to the accuracy of the data. Data that possesses all five of these attributes is what is needed for the successful training of machine learning systems.

For First Nations, there is danger both to being included in the data used to train AI systems, and danger to being excluded from these datasets. As we discussed, it is imperative for data sovereignty that First Nations have control over whether, when, and how their data is used. This includes its use in the training of AI systems. However, AI developers have a history of using data without permission, and often with little regard for privacy considerations. As examples, we can look at the training data used for large generative AI systems like chatGPT and DALL-E. Recall from chapter 2 that generative

⁷⁸ First Nations Information Governance Centre, 2019

⁷⁹ First Nations Information Governance Centre, 2020

⁸⁰ Sagioglu and Sinanc, 2013

AI systems are systems capable of generating their own content in response to a prompt, either producing a text response in the case of chatGPT or an image for DALL-E. Both systems, along with other generative AI systems, were trained on huge amounts of data that was available on the web. This included copyrighted materials, and this has led to lawsuits against the developers of both programs.⁸¹ These cases have not yet been decided, and as of now there is no precedent about whether the inclusion of copyrighted materials in a training dataset counts as copyright infringement or not. Similarly, there are worries about these systems being trained on personal data. Most companies developing generative AI do not disclose exactly what data was used in training these systems. However, it is likely that personal information, such as posts on social media sites, are part of the data used to train these systems, and this data was gathered and used without consent.

Both copyrighted information and personal information are much better protected by privacy laws than First Nations information. First Nations information included community level data that does not include personal identifiers, and thus falls outside the purview of most traditional privacy laws. Many of the traditional stories and artworks are not eligible for copyright, or even more problematically can be copyrighted by those who record them rather than by the communities whose knowledge they are.⁸² Currently is unclear to what degree privacy law and copyright apply to training data, but whatever the results turn out to be important First Nation data will remain outside the remit of these laws.

For chatGPT, the First Nation data included in its training data likely include First Nations stories and traditional knowledge where this has been made available online, as well as publicly available information about First Nations communities such as are made available by governments, in research studies, and published by organizations that serve First Nations communities. All this data was used without any engagement with First Nations themselves. For DALL-E, it is likely that the training data included images of First Nations artworks, traditional and sacred items, and other such images that fall under First Nation data.

Non-generative machine learning AI systems are also likely to make use of First Nation data. The use of machine learning is proliferating across both the public and private sectors. To take a concrete example, we can look at the healthcare sector. There is increasing interest in making use of machine learning technologies to improve healthcare delivery. Machine learning has been used in improving diagnostic accuracy,⁸³ coming up with novel medications,⁸⁴ delivering medical care personalized to

⁸¹ Appel et al., 2023

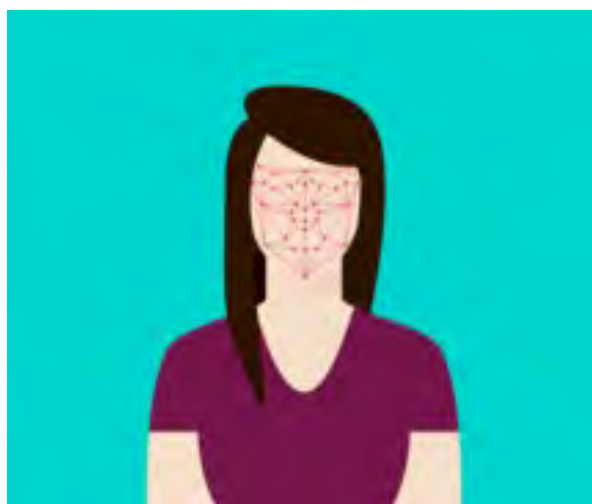
⁸² Munzer and Raustiala, 2009

⁸³ Kim H-E, et al., 2020; Alfaras et al., 2019

⁸⁴ Jumper, et al., 2021

the individual,⁸⁵ improving population health management,⁸⁶ optimizing drug dosages,⁸⁷ and more. In all these cases, the machine learning models will rely on training sets made up of sensitive health data, including data from First Nations. These models will also need to be checked for their accuracy across different historically marginalized groups, including First Nations, to avoid algorithmic bias (as discussed in chapter 4). Doing so requires access to First Nations health data to compare the accuracy of the AI system for First Nations individuals to its accuracy in the wider population. There is therefore a need to ensure that OCAP[®] is followed for all of these uses of First Nations data.

At the same time, there is a risk of First Nations data not being included in the training data for machine learning systems. While such exclusion might alleviate data sovereignty concerns, it raises the risk of algorithmic discrimination. Machine learning systems are only as accurate as the data they are trained on. If a population, such as First Nations, are excluded from the training dataset, then there is a real risk of the system working less well for that population. As an example, we can look at the case of facial recognition technology. A study looked at facial recognition software that tried to determine the gender of the person in the image, and found that while such technology was highly accurate for white men, with error rates of 0.8%, it was much worse at classifying black women, with error rates as high as 34.7%.⁸⁸ This is likely due to the overrepresentation of white men in the collections of faces used as training data for these systems, with the dataset used to test the system's performance being more than 77% male and more than 83% white.⁸⁹



Looking to the examples of machine learning being used in medicine that we considered above, we can see the significant consequences if the error rates of such systems are substantially higher for First Nations people. This could perpetuate, or exacerbate, the existing health inequalities that afflict First Nations. As such, demanding that First Nations data be removed from the training sets of AI systems comes with serious consequences of its own.

⁸⁵ Johnson et al., 2021

⁸⁶ Nelson et al., 2019

⁸⁷ Blasiak et al., 2022

⁸⁸ Buolamwini and Gebru, 2018

⁸⁹ Hardesty, 2018

To preserve data sovereignty, therefore, First Nations must be included in the datasets used to train AI systems but be included on First Nations own terms. Permission to use such data should be required, and rights holders properly consulted and given authority over how the data is used. Ideally this requirement would be laid out in legislation, rather than left up to the decision of individual developers. It is notable that the new privacy law being debated in parliament currently makes no mention of Indigenous data sovereignty and applies no specific rules around the use of First Nations data. As John Borrows and Lisa Austin point out “what has been absent is serious consultation with Indigenous communities and any attention at all to whether Bill C-27 is consistent with the federal government’s obligation to implement the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP).”⁹⁰ This is a major missed opportunity to recognize First Nations data sovereignty.

5.4 Generative AI outputs and data sovereignty

As well as worries about the data used to train AI, the outputs of generative AI can also raise issues relating to data sovereignty. As an example, consider an image-generating AI system like DALL-E. This system could be provided with a prompt to create an image based off First Nations art. The resulting image would resemble, and be based off, First Nations artistic works, but would not itself be copyrighted by any First Nation artist or group, and its generation or use would not be restricted in any way given the current laws. Similar issues could arise for text based off First Nations traditions and stories.

It is currently unclear whether there can be copyright in AI generated images or text. The Government of Canada is currently consulting on copyright in the age of Artificial Intelligence and have published a consultation paper considering these issues.⁹¹ The paper lays out three potential approaches to copyright in AI generated content. The first is to deny copyright to such works, judging them to fall immediately into the public domain. The second approach would be to follow the lead of the UK, and attribute authorship on AI-generated works to the person who arranged for the work to be created. The third approach would be to create a new set of rights for AI generated works, one that grants some economic rights on AI-generated works to a person who did not provide any original contribution to the output, such as the AI developer, deployer, or user. However, these rights would not be the full set of rights generally associated with authorship, and the person would not be deemed to be an author of the work.

None of these three approaches would provide any protection to First Nations for people creating and using images and text that resembles and is based on their traditional art and culture. The rights considered under approaches two and three are

⁹⁰ Burrows and Austin, 2022

⁹¹ ISED, 2023

only for those who participate in the creation of the work, which would be to grant rights to those non-First Nations people who use AI to replicate First Nations art. If the results are instead deemed public domain, as in the first approach, then again there would be no protection for First Nations from others making, using, and selling such images for their own profit.

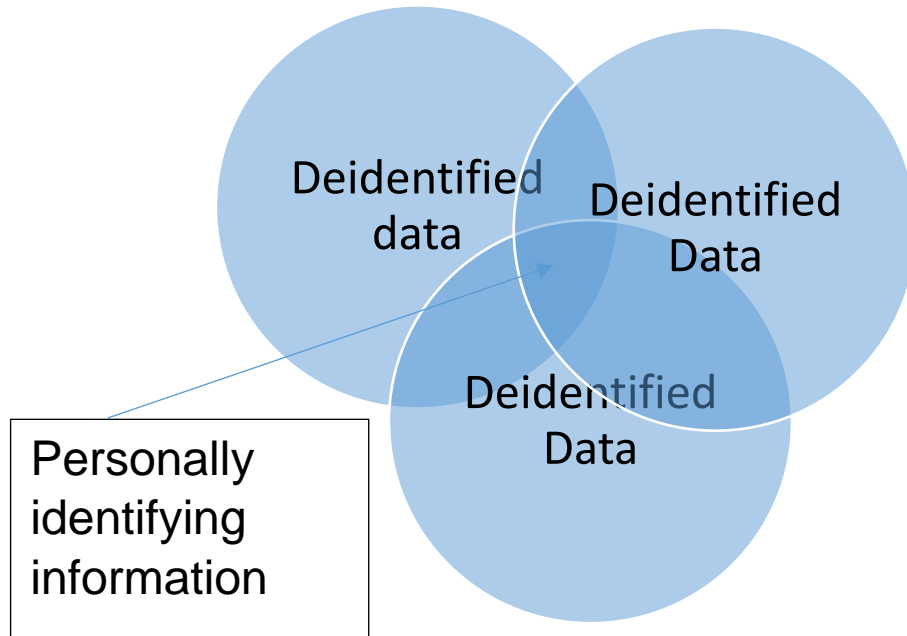
5.5 AI identifying First Nations

A separate risk to First Nations data sovereignty lies in AI's ability to identify First Nations data even when it is not explicitly labelled as such. One of the strengths of machine learning is pattern recognition; often they can pick up subtle patterns in the data that are invisible to humans. This allows machine learning to produce results that go beyond what humans can arrive at on their own. However, it also introduces new privacy risks. Machine learning can potentially discover identifying information about individuals even when that data has been explicitly removed.

For example, data is often deidentified to protect privacy. Deidentification is the process of removing personally identifying information from a dataset to make it more difficult for individuals to be identified within that dataset. However, AI raises the risk of reidentification, by combining multiple sources of data together. This is referred to as the mosaic effect, where individually harmless pieces of data combine to threaten privacy.⁹² As an example, a large dataset of New York City cab trips was released under a freedom of information request. Identifying information on the drivers was stripped from the dataset before its release. However, someone was able to correlate the timing of the trips with the timing of the five Muslim calls to prayer, identifying taxis that were always idle during these times. By so doing, the individual was able to identify which taxi drivers were likely to be Muslim.⁹³ This particular example of the mosaic effect did not involve AI. However, AI vastly increases the likelihood of this kind of reidentification because it can correlate such vast amounts of data from so many different sources. This raises the possibility of identifying First Nations populations in datasets that do not explicitly label them as such. Because of this possibility, datasets that do not directly identify First Nations individuals or communities might still raise issues of First Nations data sovereignty.

⁹² LaFever, 2023

⁹³ Franceschi-Bicchierai, 2015



In order to address this risk, First Nations will need to be able to assert their data sovereignty rights over not only datasets that contain explicit First Nations identifiers, but also those that could potentially reidentify First Nations within the data. This will require those who gather and manage such datasets to be made aware of this risk and of the principles of First Nations data sovereignty.

6. Opportunities for First Nations in AI

6.1 Introduction

A number of potential dangers from AI for First Nations have been canvassed in this paper so far. Alongside these dangers, however, there are opportunities for First Nations. AI is a powerful technology, and properly harnessed it has the potential to support the flourishing and growth of First Nations communities and traditions. The key is to identify these opportunities and take advantage of them, while resisting the rush to adopt AI in ways that put First Nations at risk.

6.2 First Nations Language Revitalization

One of the most exciting possible uses of AI for First Nations is to revitalize First Nations languages. There are over 70 Indigenous languages spoken in Canada. However, many of them are at risk, with falling number of both speakers and of those for whom an Indigenous language is their first language. The number of people who can speak an Indigenous language well enough to carry on a conversation has declined by 4.3% from 2016.⁹⁴

One potential use for AI is to create systems that can help to teach more First Nations people to speak their languages and allow people to interact with technology in their own language. A first step in this work is to create AI systems that can recognize First Nations languages, in the same way that voice recognition systems such as Apple's Siri or Google's Alexa virtual assistants can recognize English and other languages. However, training AI systems to recognize languages requires huge amounts of training data. This is no problem for English, where native speakers have created huge volumes of recorded speech that can be used for AI training. However, many First Nations languages do not have the same volume of examples of recorded speech and may have relatively few speakers from which to generate additional data.

The First Language AI Reality, or FLAIR, project is seeking to address this problem. Housed at MILA, a prestigious AI research institute based in Montreal, FLAIR is headed by Micheal Running Wolf and Caroline Running Wolf. The aim of the project is to allow for the creation of automatic speech recognition systems for indigenous languages without requiring the huge quantities of training data that have traditionally been required. This project is currently focusing on the Wakashan language family spanning British Columbia, Canada and Washington State, USA, with the long-term goal of expanding to other Indigenous languages in North America, such as Algonquin languages in the Northeast, eventually bringing Voice AI to Indigenous communities worldwide.⁹⁵

The long-term vision of the project is even more expansive. They paint the following picture:

⁹⁴ Statistics Canada, 2023

⁹⁵ MILA, n.d

“Imagine an inclusive Metaverse where Indigenous youth across North America reconnect with their heritage. In this Metaverse Lakota hunters use their mother tongue to coordinate a bison hunt in the Great Plains. To the west, Makah canoes cross the Salish Sea to a community reunion with Kwakwaka'wakw speakers and ask permission to come ashore on their Canoe Journey. We are convinced that Voice AI will be intrinsic to the Metaverse experience which in turn can facilitate intergenerational language transmission for Indigenous languages.”⁹⁶

This is a picture of AI technology being used to revitalize First Nations languages and culture and connect First Nations people to one another and to their nations. While the full realization of this vision may still be a long way off, it shows how we can reimagine AI technology as an opportunity, not just a threat, to First Nations peoples.

6.3 Support for First Nation Governments

First Nations communities have the right to self-government and self-determination. This is an inherent right of First Nations peoples and is recognized by the federal government both as a right under section 35 of the constitution and in their adoption of the UNDRIP act, which commits Canada to the principles set out in UNDRIP. While the government's actions do not always reflect these commitments, the commitment itself is clear.

However, despite this inherent right, many aspects of self-government can create challenges for First Nations. Many First Nations are quite small, and the required administration for providing services and exercising self-determination can be a challenge. Much of this challenge stems from under-funding by the government of First Nations. This needs to be addressed, and the advent of AI systems does not change this fact. However, the use of AI systems might allow for increased capacity for First Nations communities to self-govern and reduce reliance on the federal and provincial governments.

AI systems can, if properly used, ease the administrative burden of self-government. While care must be taken to ensure fairness and to allow for a right to object to AI decision making, as discussed above in section 4, AI systems can potentially automate time-consuming tasks and free up time for leadership and staff of First Nations. As an example, some First Nations currently struggle to administer their own Ontario Works programs and must rely on nearby municipalities to administer the program on their behalf. If an AI system could automate some of the work of administering the program, then more communities could take advantage of this and avoid having to rely on non-First Nations governments.

6.4 Preserving the Land

AI technology poses significant environmental risks. The training of large language models (LLM's), such as GPT-3, requires huge amounts of energy. It is estimated that

⁹⁶ Mila, n.d.

training a large model can release as much as 284 tons of carbon dioxide, contributing to global warming.⁹⁷ In comparison, the average person causes the release of 4.7 tons of carbon dioxide per year.⁹⁸ The current rush to create ever larger machine learning models therefore creates a real risk of environmental damage.

However, there are also opportunities to use AI to protect the environment in collaboration with Indigenous peoples who protect the land. As an example, we can look to the PolArctic project conducted in Sanikiluaq, an Inuit community in Nunavut in 2021. The project used AI to identify areas of the sea near Sanikiluaq that were likely to contain higher concentrations of scallops, clams and kelp, providing a boon to the local mariculture industry. Crucially, the project incorporated and relied on the knowledge of the local Inuit peoples, incorporating their traditional knowledge alongside other sources of information to inform the AI model. The project was described as “the first AI model of its kind to treat Indigenous Knowledge and western science as equals, training with and validating both knowledge systems.”⁹⁹ The knowledge generated by the AI model could also be used to choices about sewage disposal, creating shipping routes, and protecting habitat areas.

Another example of AI used to protect the land is the Temb  tribe in Northern Brazil. They collaborated with a non-profit organization called Rainforest Connection to create a low-cost alert system to monitor deforestation. The group used an open-source AI software called TensorFlow to identify the sounds of chainsaws and logging from the background noise of the Amazon rainforest. Text alerts would then be sent to Temb  patrols, who could intervene to protect the forest.¹⁰⁰

There are opportunities, then, to leverage AI in partnership with First Nations to help protect the land. It is important that First Nations traditional knowledge be respected and protected in these projects, and that the data is used in accordance with the principles of data sovereignty discussed in section 5 above. If First Nations are treated as genuine partners in these endeavors, then AI can offer new opportunities to help First Nations people protect the land as they have done for thousands of years.

⁹⁷ Bender, Gebru, et al., 2021. Notably, following the release of this paper, which criticized large language models for their climate impact and the risk of bias, one of the authors, Temnit Gebru, was fired from Google’s ethical AI team.

⁹⁸ IEA, 2023

⁹⁹ Canavera, 2022

¹⁰⁰ Dimock, 2022

Glossary

Algorithm: A set of rules for a computer program to follow to arrive at an output.

Algorithmic Bias: Instances where an artificial intelligence system produces outputs that are biased against certain groups, especially historically marginalized groups.

Artificial Intelligence (AI): Computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Artificial General Intelligence: AI systems that are capable of doing everything a human reasoner can do, and potentially more. In other words, an AI that can function as well as, or better than, a human across all domains requiring intelligence. This kind of AI does not currently exist.

Artificial Narrow Intelligence: AI systems that are able to do tasks that traditionally have required human intelligence, but only for specific narrow areas of application. For example, an AI facial recognition system can recognize individuals from their faces but has no broader capacity to reason. All currently existing AI systems are of this type.

Artificial Neural Networks: A method of machine learning which is inspired by the structure of neurons in the human brain. It features a number of nodes connected to one another in layers. There is an input layer, into which data is fed. The data is then processed through a number of hidden layers, and the output is finally provided by the output layer.

Artificial Intelligence Model: A program that applies one or more algorithms to data to recognize patterns, make predictions or make decisions without human intervention.

Black-box AI: An AI system where humans cannot explain what factors lead an AI system to produce the outputs it arrives at. Many machine learning AI systems are black boxes.

Data: Observations or measurements (unprocessed or processed) represented as text, numbers, or multimedia.

Data Scraping: A technique in which a computer program extracts data from output generated from another program. One variety of data scraping would be for a program to extract data from a website or multiple websites. This is often used as a means of acquiring training data for training machine learning systems.

Data Sovereignty: The right to control what data is collected, who has access to it, and how it is managed.

Deep Learning: A type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data.

De-identification: A process used to protect the privacy of individuals by removing information that identifies an individual or for which there is a reasonable expectation that it could be used, either alone or with other information, to identify an individual.

Generative AI: A type of AI system which is capable of generating its own content, such as writing, images, or video, based on a prompt describing the desired output.

Machine Learning: Computer systems that are able to learn and adapt independent of human instructions by identifying patterns in training data.

Mosaic Effect: When bringing together multiple datasets allows for significant new information to be revealed. This can threaten privacy, when personally identifying information is inferred from datasets that have been deidentified.

OCAP®: An acronym for ownership, control, access, and possession. It is a data sovereignty principle that dictates the rights of First Nations over their data.

Personally Identifying Information: Information that can, either on its own or when combined with other relevant data, can identify an individual.

Proxy: A proxy is a variable that is meant to track the desired outcome and is used when the desired outcome cannot be measured directly. For example, scores on standardized tests might be used as a proxy for educational success.

Symbolic Artificial Intelligence: An AI system developed by programming a computer with a series of symbols along with rules and axioms for the manipulation of these symbols.

Training Data: The initial data that is used to train a machine learning system. Machine learning is based on identifying patterns in the training data, so biased or incomplete training data can lead to discriminatory AI systems.

References

Alfaras M., Soriano MC., Ortín S. (2019) “A fast machine learning model for ECG-based Heartbeat classification and arrhythmia detection.” *Front Phys.*7.

<https://doi.org/10.3389/fphy.2019.00103>.

Amdur, E. (2023) “Venture Capital In AI – Where And How Much” *Forbes*
<https://www.forbes.com/sites/eliamdur/2023/11/16/venture-capital-in-ai--where-and-how-much/>. Last accessed February 21st, 2024

Angwin, J., Larson, J. Mattu, S. and Kirchner, L. (2016) “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica*, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>, accessed March 1st, 2024

Angwin, J., and Parris, T. (2016) “Facebook Lets Advertisers Exclude Users by Race.” *ProPublica*, <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race> accessed February 14th 2024

Anyoha, R. (2017) “The History of Artificial Intelligence”, *Harvard*
<https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/> accessed March 1st, 2024

Appel, G., Neelbauer, J., Schweidel, D. (2023) “Generative AI Has an Intellectual Property Problem.” *Harvard Business Review*. <https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem>

Arif, S. (2023) “An Overview of the Rise and Fall of Expert Systems.” *Medium*.
<https://medium.com/version-1/an-overview-of-the-rise-and-fall-of-expert-systems-14e26005e70e> accessed March 1st 2024

Assembly of First Nations (2023) “Bill C-27: The Digital Charter Act, 2023 and First Nations Rights”,
<https://www.ourcommons.ca/Content/Committee/441/INDU/Brief/BR12885140/br-external/AssemblyOfFirstNations-e.pdf> accessed February 27th 2024

Babu, D., (2023) “Reward Hacking in Large Language Models (LLMs)” 2023.
<https://medium.com/@prdeepak.babu/reward-hacking-in-large-language-models-llms-c57abbc0cde7>, accessed November 6th 2023

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). “On the dangers of stochastic parrots: Can language models be too big?”. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623)

Biden, J. (2023) “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence” E.O. 14110 of 2023-10-30

Bill C-27 (2021) “An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts” First Session, Forty-fourth Parliament

Blasiak A, Truong A, Jeit W, Tan L, Kumar KS, Tan SB, et al. (2022) “PRECISE CURATE.AI: a prospective feasibility trial to dynamically modulate personalized chemotherapy dose with artificial intelligence.” *J Clin Oncol*.40(16suppl):1574–4. https://doi.org/10.1200/JCO.2022.40.16_suppl.1574.

Buolamwini, J., & Gebru, T. (2018) “Gender shades: Intersectional accuracy disparities in commercial gender classification.” In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Buolamwini, J. (2023). “Unmasking AI: my mission to protect what is human in a world of machines.” Random House.

Burrows, J. and Austin, L. (2022) “The Digital Charter Implementation Act ignores Indigenous Data Sovereignty”, Schwartz Reisman Institute for Technology and Society, <https://srinstitute.utoronto.ca/news/digital-charter-implementation-act-ignores-indigenous-data-sovereignty>, accessed February 27th 2024

Cadwalladr, C. (2016) “Google, democracy and the truth about internet search” *The Guardian* <https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook>, accessed October 16th 2023.

Canavera, L. (2022) “Blending Indigenous Knowledge and artificial intelligence to enable adaptation”. World Wildlife Federation. <https://www.arcticwwf.org/the-circle/stories/blending-indigenous-knowledge-and-artificial-intelligence-to-enable-adaptation/>. Accessed on February 21st, 2024.

Champagne, F. (2023) “Correspondence from the Honourable François-Philippe Champagne, Minister of Innovation, Science and Industry - Amendments to AIDA - 2023-11-28.” <https://www.ourcommons.ca/DocumentViewer/en/44-1/INDU/related-document/12751351> accessed March 1st 2024

Chandran, R., Smith, A., Ramos, M. (2023) “AI boom is dream and nightmare for workers in Global South.” *Context*, Thomas Reuters Foundation <https://www.context.news/ai/ai-boom-is-dream-and-nightmare-for-workers-in-global-south>. Accessed October 16th, 2023.

Clark, J., and Amodei, D. (2016) “Faulty reward functions in the wild.” OpenAI. <https://openai.com/research/faulty-reward-functions> accessed March 1st, 2024

Dai, J., & Brown, S. M. (2020). Label bias, label shift: Fair machine learning with unreliable labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments* (Vol. 12).

Dastin, J. (2018) "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters <https://www.reuters.com/article/amazoncom-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1VB1FQ/?feedType=RSS%26feedName=companyNews> Accessed November 6th, 2023

DataCamp (2023) "What is Symbolic AI?" <https://www.datacamp.com/blog/what-is-symbolic-ai>, accessed September 22nd, 2023

Department of Science, Innovation and Technology (2023) "A pro-innovation approach to AI regulation." <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper> accessed February 27th, 2024

Dimmock, W. C. (2022) "AI Can Help Indigenous People Protect Biodiversity", *Scientific American*, <https://www.scientificamerican.com/article/ai-can-help-indigenous-people-protect-biodiversity/>. Accessed on February 21st, 2024

Duarte, F. (2023) "Amount of Data Created Daily (2024)" *Exploding Topics* <https://explodingtopics.com/blog/data-generated-per-day> accessed accessed March 1st 2024

Dyvik, E. (2023) "Biggest companies in the world by market value 2023." *Statista*. <https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/> accessed March 1st, 2024

First Nations Information Governance Centre. (2016). Pathways to first nations' data and information sovereignty. In T. Kukutai, & J. Taylor (Eds.), *Indigenous Data Sovereignty: Toward an Agenda* (pp. 139–155). Australian National University Press.

First Nations Information Governance Centre. (2019). "The First Nations principles of OCAP®." <https://fnigc.ca/wp-content/uploads/2021/08/OCAP-Brochure-2019.pdf> Accessed May 2nd, 2022

First Nations Information Governance Centre. (2020). *First Nations Data Governance Strategy*. Accessed March 9th 2022, https://fnigc.ca/wp-content/uploads/2020/09/FNIGC_FNDGS_report_EN_FINAL.pdf

Foote, D. (2021) "A Brief History of Machine Learning." *Dataversity*. <https://www.dataversity.net/a-brief-history-of-machine-learning/> accessed March 1st 2024

Franceschi-Bicchierai, L. (2015) “Redditor cracks anonymous data trove to pinpoint Muslim cab drivers” Mashable <https://mashable.com/archive/redditor-muslim-cab-drivers>, accessed December 7th 2023

Google (2023) “Google Diversity Annual Report 2023” https://static.googleusercontent.com/media/about.google/en//belonging/diversity-annual-report/2023/static/pdfs/google_2023_diversity_annual_report.pdf?cachebust=2943cac (last accessed February 14th, 2023)

Gurney, K. (2018). An introduction to neural networks. CRC press.

Hardesty, J. (2018) “Study finds gender and skin-type bias in commercial artificial-intelligence systems” MIT news <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>, accessed December 7th 2023

Heaven, W. D. (2022) “Why Meta’s latest large language model survived only three days online”, Technology Review <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>, accessed November 6th 2023

Hughes, A. (2023) “ChatGPT: Everything you need to know about OpenAI’s GPT-4 tool” Science Focus <https://www.sciencefocus.com/future-technology/gpt-3>, accessed March 1st 2024

IEA. (2023) “The world’s top 1% of emitters produce over 1000 times more CO2 than the bottom 1%.” IEA, Paris. <https://www.iea.org/commentaries/the-world-s-top-1-of-emitters-produce-over-1000-times-more-co2-than-the-bottom->, accessed February 21st, 2024

ISED (2023) “The Artificial Intelligence and Data Act (AIDA) – Companion document” <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> accessed March 1st 2024

ISED (2023) “Consultation on copyright in the age of generative artificial intelligence.” <https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-age-generative-artificial-intelligence#s1> accessed March 1st 2024

Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. (2021) “Precision Medicine, AI, and the future of Personalized Health Care.” *Clin Transl Sci.*;14(1):86–93. <https://doi.org/10.1111/cts.12884>.

Jumper, J., Evans, R., Pritzel, A. et al. (2021) “Highly accurate protein structure prediction with AlphaFold.” *Nature* 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>

Kim H-E., Kim H.H., Han B-K., Kim K.H., Han K., Nam H., et al. (2020) “Changes in cancer detection and false-positive recall in mammography using Artificial Intelligence: a retrospective” Multireader Study. *Lancet Digit Health*. 2(3).

[https://doi.org/10.1016/s2589-7500\(20\)30003-0](https://doi.org/10.1016/s2589-7500(20)30003-0).

LaFever, G. (2023) “Beyond GDPR: Unauthorized reidentification and the Mosaic Effect in the EU AI Act.” International Associations of Privacy Professionals.

<https://iapp.org/news/a/beyond-gdpr-unauthorized-reidentification-and-the-mosaic-effect-in-the-eu-ai-act/> accessed March 1st, 2024

Li S., Zhao R., Zou H. (2021) “Artificial intelligence for diabetic retinopathy.” *Chin Med J (Engl)*.135(3):253–60. <https://doi.org/10.1097/CM9.0000000000001816>.

Macukow, B. (2016). Neural Networks – State of Art, Brief History, Basic Models and Architecture. In: Saeed, K., Homenda, W. (eds) *Computer Information Systems and Industrial Management. CISIM 2016. Lecture Notes in Computer Science()*, vol 9842. Springer, Cham. https://doi.org/10.1007/978-3-319-45378-1_1

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J.C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. (2023) “The AI Index 2023 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

McInerney v MacDonald (1992) CanLII 57 (SCC), 2 SCR 138, online:

<http://canlii.ca/t/1fsbl>.

Metz, C. (2016) “In two moves, AlphaGo and Lee Sedol redefined the future”, *Wired*.

<https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/> accessed March 1st, 2024

MILA (n.d.) “First Languages AI Reality (FLAIR) Initiative”

<https://mila.quebec/en/project/flair-initiative/> accessed on March 1st 2024

Munzer, S. R., & Raustiala, K. (2009). “The uneasy case for intellectual property rights in traditional knowledge.” *Cardozo Arts & Ent. LJ*, 27, 37.

National Institute of Standards and Technology (2023) “Artificial Intelligence Risk Management Framework” <https://doi.org/10.6028/NIST.AI.100-1> accessed March 1st 2024

Nelson KM, Chang ET, Zulman DM, Rubenstein LV, Kirkland FD, Fihn SD. (2019) “Using Predictive Analytics to Guide Patient Care and Research in a National Health System.” *J Gen Intern Med*.34(8):1379–80. <https://doi.org/10.1007/s11606-019-04961-4>.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". *Science*, 366(6464), 447-453.

OECD (2019) "Recommendation of the Council on Artificial Intelligence"

<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

Oxford Dictionary of Phrase and Fable (2 ed.) (2005) "Artificial Intelligence." Oxford University Press

<https://www.oxfordreference.com/display/10.1093/acref/9780198609810.001.0001/acref-9780198609810-e-423> accessed March 1st, 2024

Pocock, K. (2023) "What Was Dall-E 2 Trained On?" PC Guide

<https://www.pcguides.com/apps/what-was-dall-e-2-trained-on/> accessed March 1st 2024

Poursabzi-Sangdeh, F., Goldstein, D., Hofman, J., Wortman Vaughan, J., Wallach, H. (2021) "Manipulating and measuring model interpretability." Proceedings of the 2021 CHI conference on human factors in computing systems.

Rainie SC, Rodriguez-Lonebear D, & Martinez A. (2017). Policy Brief: Data Governance for Native Nation Rebuilding. (Version 2). Tucson: Native Nations Institute, The University of Arizona, 2022,

https://climas.arizona.edu/sites/climas.arizona.edu/files/Policy_Brief_Data_Governance_for_Native_Nation_Rebuilding_Version_2.pdf Accessed March 1st 2024

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 (2021) <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206> Accessed March 1st 2024

Ribeiro, M. T., Singh, S., Guestrin, G (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", arXiv:1602.04938,

<https://doi.org/10.48550/arXiv.1602.04938>

Sagiroglu, S., & Sinanc, D. (2013, May). "Big data: A review." In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.

Scassa, T. (2018) "Data Ownership" CIGI Papers No. 187, Ottawa Faculty of Law Working Paper No. 2018-26, Available at SSRN: <https://ssrn.com/abstract=3251542> or <http://dx.doi.org/10.2139/ssrn.3251542>

Scassa, T. (2023a) "High-Impact AI Under AIDA's Proposed Amendments (Part II of a Series)", personal blog

https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=375:high-impact-ai-under-aidas-proposed-amendments-part-ii-of-a-series&Itemid=80, accessed February 27th 2024

Scassa, T. (2023b). "Oversight and Enforcement in the AIDA Amendments (Part III of a series)", personal blog

https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=376:oversight-and-enforcement-in-the-aida-amendments-part-iii-of-a-series&Itemid=80 Accessed March 1st, 2024

Scassa, T. (2023c) "Comparing the UK's proposal for AI governance to Canada's AI bill" Personal blog.

http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=370:comparing-the-uks-proposal-for-ai-governance-to-canadas-ai-bill&Itemid=80 accessed March 2st 2024

Schuchmann, Sebastian. (2019). "Analyzing the Prospect of an Approaching AI Winter." DOI: 10.13140/RG.2.2.10932.91524.

Roser, M. (2022) "The brief history of artificial intelligence: The world has changed fast – what might be next?" OurWorldInData.org. 'https://ourworldindata.org/brief-history-of-ai' accessed March 1st, 2024

Statistics Canada (2023) "Indigenous languages across Canada", <https://www12.statcan.gc.ca/census-recensement/2021/as-sa/98-200-X/2021012/98-200-X2021012-eng.cfm> (last accessed February 16th 2024)

Tangalakis-Lippert, K. (2023) "Clearview AI scraped 30 billion images from Facebook and other social media sites and gave them to cops: it puts everyone into a 'perpetual police line-up'" Business Insider <https://www.businessinsider.com/clearview-scraped-30-billion-images-facebook-police-facial-recognition-database-2023-4>, accessed December 1st 2023

Thompson, N., Ge, S., Manso, G. (2022) "The Importance of (Exponentially More) Computing Power." arXiv:2206.14007, <https://doi.org/10.48550/arXiv.2206.14007>

Tromp, J. (2016). "The Number of Legal Go Positions." In: Plaat, A., Kusters, W., van den Herik, J. (eds) Computers and Games. CG 2016. Lecture Notes in Computer Science(), vol 10068. Springer, Cham. https://doi.org/10.1007/978-3-319-50935-8_17

Turing, Alan M. (1950) "Computing machinery and intelligence." In The Essential Turing: the Ideas That Gave Birth to the Computer Age (2012): 433-464.

United Nations (General Assembly). (2007). Declaration on the Rights of Indigenous People.

Vincent, J. (2018) "Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women." The Verge

<https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report>
accessed February 13th 2024

White House Office of Science and Technology Policy (2022) "Blueprint for an AI bill of rights: making automated systems work for the American people"

<https://www.whitehouse.gov/ostp/ai-bill-of-rights> accessed March 1st 2024