A Code of Conduct for the Ethical Use of Artificial Intelligence in Canadian Financial Services

Smith School of Business, Queen's University
Stephanie Kelley
Dr. Yuri Levin
Dr. David Saunders

Draft Version 1.11
Last Updated: October 19th, 2018

Table of Contents

1. Introduction

1.1. *What is the Code of Conduct for the Ethical Use of Artificial Intelligence in Canadian Financial Services?*

The Code of Conduct (herein referred to as "the Code") is a set of principles developed to guide Canadian financial services organizations in the ethical use of artificial intelligence (AI). The rapid technical advances in AI have driven significant innovation but have also created a gap between the technology and the ethical guidelines and regulations for its application. The Code is designed to close this gap and provide practical guidance for Canadian financial services organizations to navigate the day-to-day ethical implications that arise when using AI.

1.2 Scope

The Code is designed to offer guidance to all organizations that provide financial services, use artificial intelligence, and operate in Canada. It is designed to be an addition to existing legal, privacy, compliance, regulatory, ethical and other related guidelines that guide the actions of financial services organizations in Canada. Organizations that choose to abide by the Code should ensure their internal governance frameworks enable behaviour and actions in line with the principles set out below in the Summary of Principles section. In addition, they may also choose to define their own principles and ethical values in line with the summary principles presented.

1.3 Key Definitions

Artificial intelligence (AI): The ability of a machine to perform tasks commonly associated with intelligent human behaviours including, but not limited to learning from, and acting on information[1,2].

AI Application: Artificial intelligence that is designed to complete a specific task or process.

Autonomy: The ability of an AI application to perform tasks in response to environmental inputs, independently of real-time human input[3]. Autonomy can be assigned to an AI application in varying degrees, ranging from human-controlled, where a machine requires real-time human input to perform a task, to fully autonomous, where a machine performs tasks with no real-time human input. In between these two extremes are semi-autonomous AI applications which can perform some, but not all tasks, independently of real-time human input.

AI Approver: An individual, or group of individuals, who approve an AI application prior to production.

AI Designer: An individual, or group of individuals, who provide the technical input used to generate an AI application.

[1] Artificial Intelligence | Definition of Artificial Intelligence by Merriam-Webster. (2018). Retrieved September 13, 2018, from https://www.merriam-webster.com/dictionary/artificial intelligence

[2] Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., … Trench, M. (2017). *Artificial Intelligence - The Next Digital Frontier*. https://doi.org/10.1016/S1353-4858(17)30039-9

[3] Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, *18*(2), 149–156. https://doi.org/10.1007/s10676-016-9400-6

AI User: An individual, or group of individuals, who utilize an AI application and its output.

AI Owner: An individual, or group of individuals, who request the creation of an AI application; they may also be the AI user.

Individuals: Individuals who are or could be affected by the output of an AI application.

AI Stakeholder: An individual, or group of individuals, who have a vested interest in an AI application. This includes, but is not limited to the AI approver, AI designer, AI user, AI owner, and individuals.

Ethical Implication: A problem, situation, or opportunity that requires a decision maker to choose among several actions that must be evaluated as right or wrong, ethical or unethical; however, by nature involves conflicts of values and often requires a degree of ethical judgement rather than a binary evaluation[4].

Material: A property of an AI application, a decision made by an AI application, or a decision made using AI application output that indicates it has a significant or relevant impact on AI stakeholders, especially individuals.

Organization: Any firm operating in Canada that provides financial services and utilizes AI.

[4] Khalil, O. E. M. (1993). Artificial Decision-Making and Artificial Ethics: A Management Concern. *Journal of Business Ethics*, *12*(4).

2. Summary of Principles

2.1 Principles of Fairness

Bias and Discrimination

1. Organizations should not unfairly discriminate against nor intentionally or unintentionally disadvantage individuals.
2. Organizations should conduct an ethics review of all material AI applications prior to their production to ensure unfair discrimination, and intentional and unintentional disadvantage are avoided.
3. Organizations should have in place an appropriate internal governance framework to determine what constitutes unfair discrimination.

Justifiability

4. Organizations should understand how the aggregate input data for a material AI application impacts the output to ensure unfair discrimination, and intentional and unintentional disadvantage are avoided.
5. Organizations should ensure the use of all data attributes adhere to the laws, ethical standards, codes, and values that govern their organization and industry.

2.2 Principles of Accountability

Responsibility & Accountability for AI Application Output

6. Organizations are responsible and accountable for the output of their AI applications, whether they are designed internally or externally.
7. Organizations must ensure responsibility and accountability are assigned to specific individuals for the output of each AI application in production, in line with the application's materiality.
8. The AI approver must approve all AI applications prior to production, and any changes made after the initial approval.

Autonomy

9. Organizations are accountable for the level of autonomy assigned to an AI application and should ensure autonomous decisions are aligned with the AI owner, AI user, AI designer, and AI approver, with consideration of the impact on individuals and the resulting ethical implications.
10. Decisions made by an AI application and decisions made with AI application output should be held to the same ethical standards as decisions made by humans, even if the AI application has been assigned full autonomy.
11. Decisions made by an AI application and decisions made with AI application output can significantly impact the workforce; organizations are expected to proactively manage these implications in line with the laws, ethical standards, codes, and values that govern their organization and industry.

Consistency

12. Organizations should ensure they have an AI model development process in place, which at minimum should include an initial validation stage, an approval stage, and ongoing monitoring to ensure alignment with the ethical principles.
13. Organizations should review all AI applications to ensure they are being used and performing as intended; they should align on the appropriate review period, which may vary depending on the materiality of the AI application.

2.3 Principles of Transparency

Explainability

14. Individuals should have access, upon request, to a reasonable explanation for any material decisions made by an AI application, or with AI application output. A reasonable explanation should at minimum include what data attributes are used, and how those data attributes impact the AI application output.

Data Privacy & Informed Consent

15. All AI applications should adhere to the data privacy laws, standards, and other related guidelines that govern the organization and industry.
16. Organizations should ensure individuals, internal and external, provide meaningful informed consent for the use of their data in circumstances where privacy is expected, per applicable local laws.

Announcing the Use of AI Applications

17. Organizations should proactively announce the use of AI applications to individuals when they are interacting directly with an AI application, and when a material decision is made by an AI application or with AI application output.

3. Principles of Fairness

Bias and Discrimination

1. Organizations should not unfairly discriminate against nor intentionally or unintentionally disadvantage individuals.
2. Organizations should conduct an ethics review of all material AI applications prior to their production to ensure unfair discrimination, and intentional and unintentional disadvantage are avoided.
3. Organizations should have in place an appropriate internal governance framework to determine what constitutes unfair discrimination.

3.1. Bias in data, algorithmic bias, and AI designer bias can lead to unfair discrimination, and intentional or unintentional disadvantage of individuals. Organizations should proactively manage all forms of bias to ensure individuals are not intentionally disadvantaged or unfairly discriminated against.

3.2. There may be instances where an organization uses certain data attributes to differentiate between individuals but doing so must never lead to intentional disadvantage or unfair discrimination.

3.3. It is prohibited, per the Canadian Human Rights Act to unfairly discriminate based on race, national or ethnic origin, colour, religion, age, sex, sexual orientation, gender identity or expression, marital status, family status, genetic characteristics, disability, or conviction for an offence for which a pardon has been granted or in respect of which a record suspension has been ordered[5].

3.4. The use of data attributes as predictors is core to many business models, but there is often no clear line between fair versus unfair discrimination. Organizations should ensure they have an appropriate internal governance framework in place to align with AI stakeholders on a definition of unfair discrimination. This definition must at minimum include the prohibited data attributes per the Canadian Human Rights Act, per Section 3.3, but may include additional data attributes based on an organization's ethical code of conduct or values.

3.5. Unfair discrimination can occur in an AI application's output despite an organization's best efforts to avoid it; organizations should therefore ensure an ethics review occurs, lead by the AI approver, to review the AI application output. Organizations should ensure the ethics review process is multi-disciplinary and should align on the format with the relevant AI stakeholders.

[5] Minister of Justice. (2017). Canadian Human Rights Act.

3.6. Example: A data scientist (the AI designer) and a risk management manager (AI owner and AI user) work together to create an AI application to automate the credit adjudication process. The AI application is designed to exclude certain data attributes that are prohibited by law and could lead to unfair discrimination of individuals. The AI application might, without the data scientist's or risk management manager's knowledge, select another data attribute to include in the model that is highly correlated with one of the prohibited data attributes. This causes the AI application output to unfairly discriminate against certain individuals based on the prohibited data attribute, preventing them from obtaining credit, despite the fact the prohibited data attribute was not included in the application. This unfair discrimination would likely only be caught through an ethics review of the AI application's output, not from reviewing the data or technical aspects of the AI application by themselves.

Justifiability

4. Organizations should understand how the aggregate input data for a material AI application impacts the output to ensure unfair discrimination, and intentional and unintentional disadvantage are avoided.
5. Organizations should ensure the use of all data attributes adhere to the laws, ethical standards, codes, and values that govern their organization and industry.

3.7. To prevent ethical implications, organizations should proactively ensure they understand the impact of the aggregate input data on the output of material AI application to ensure unfair discrimination and intentional and unintentional disadvantage are avoided. In all instances, organizations should be prepared to explain to AI stakeholders why the data attributes were chosen to include in the AI application, and how the aggregate input data impacts the AI application output.

3.8. Example: An HR analyst (the AI owner and user) works with a data scientist (the AI designer) to develop an AI application to determine the drivers of employee attrition. The HR analyst is interested in aggregate predictors of attrition and she works with the data scientist during the AI application development to protect employee privacy by excluding personally identifiable information, such as name, address, SIN, etc. The AI application output provides them with five strong predictors of attrition. The HR analyst and data scientist present the findings during the ethics review process, along with an explanation of each data attribute and its impact on the application output. During discussions with other AI stakeholders in the ethics review process they realize one of the five predictors is highly correlated with marital status, a prohibited attribute. The AI stakeholders agree that the AI application could lead to unfair discrimination against individuals, and that the correlated attribute should be removed from the AI application before it can be approved for production.

4. Principles of Accountability

Responsibility & Accountability for AI Application Output

6. Organizations are responsible and accountable for the output of their AI applications, whether they are designed internally or externally.
7. Organizations must ensure responsibility and accountability are assigned to specific individuals for the output of each AI application in production, in line with the application's materiality.
8. The AI approver must approve all AI applications prior to production, and any changes made after the initial approval.

4.1. AI applications can create unexpected and unexplainable outputs; however, the organization is responsible and accountable for all output of AI applications in production, even in instances where an AI application is provided full autonomy or designed externally.

4.2. Organizations must ensure responsibility and accountability are assigned to specific individuals for the output of each AI application to ensure ethical implications are dealt with and prevented in the future. The individuals assigned responsibility and/or accountability must have a reasonable level of understanding of the AI application. Even in the case of highly complex AI applications, organizations need to ensure clear responsibility and accountability are assigned; alternatively, the AI application should not be put into production.

4.3. Example: If an autonomous algorithmic trading AI application routes orders without an organization's knowledge causing many individuals to lose funds, who in the organization is responsible, and accountable for the mistake? Prior to the AI application's approval, the organization should have assigned clear responsibility and accountability for the trading application's output, ensuring all parties had a reasonable level of understanding of the AI application. For instance, the organization could have assigned joint responsibility to a group of data scientists (AI designers) and traders (AI users); however, accountability might have been assigned to the senior trader (AI owner) leading the project.

Autonomy

9. Organizations are accountable for the level of autonomy assigned to an AI application and should ensure autonomous decisions are aligned with the AI owner, AI user, AI designer, and AI approver, with consideration of the impact on individuals and the resulting ethical implications.
10. Decisions made by an AI application and decisions made with AI application output should be held to the same ethical standards as decisions made by humans, even if the AI application has been assigned full autonomy.
11. Decisions made by an AI application and decisions made with AI application output can significantly impact the workforce; organizations are expected to proactively manage these implications in line with the laws, ethical standards, codes, and values that govern their organization and industry.

4.4. AI applications are designed and used by humans, and it is humans who decide the degree of autonomy assigned to an AI application, whether that be human-controlled, semi-autonomous, or fully autonomous. The delegation of autonomy comes with great responsibility, and organizations would do well to remember that technology is not a perfect substitute for humans.

4.5. The degree of autonomy assigned to an AI application can have ethical implications for job displacement or job loss. Organizations should proactively manage the impact of AI applications on their workforce. Depending on the speed of AI application implementation this may require significant investment in uptraining or retraining.

4.6. Example: An organization wants to find an AI application to help them reduce the cost of providing customer service to retail banking customers. They have held a competition internally and have been presented two options, one AI application is a fully autonomous AI chat-bot that can handle 300 frequently asked questions and directly interact with customers, the other is a semi-autonomous AI chat-bot that can help a human customer service representative by recommending answers to 500 frequently asked questions but can’t interact directly with customers. The organization must choose, based on input from AI stakeholders, which system to choose, ensuring they consider the system’s degree of autonomy, how it will impact their workforce, and how they will proactively manage the ethical implications of their choice.

Consistency

12. Organizations should ensure they have an AI model development process in place, which at minimum should include an initial validation stage, an approval stage, and ongoing monitoring to ensure alignment with the ethical principles.
13. Organizations should review all AI applications to ensure they are being used and performing as intended; they should align on the appropriate review period, which may vary depending on the materiality of the AI application.

4.7. AI stakeholder input is an important mechanism to manage the ethical implications of an AI application. Organizations should therefore have a consistent model development process in place across business units to ensure they gather AI stakeholder input to help manage ethical implications.

4.8. Many AI applications are highly dynamic; they gather, process, learn from, and act on vast amounts of data, which itself is evolving over time, ultimately impacting the AI application output. As such, organizations should have in place a safe testing environment for their AI applications where individuals cannot be negatively impacted, and they should monitor the AI applications in production to ensure they are performing as intended. Due to the variety of AI applications in use, it is up to the organization to align on the appropriate review period for each AI application. Material applications that are highly dynamic may warrant more frequent review.

4.9. Example: A model development process could include seven stages that are managed iteratively: purpose, data collection, model development, validation, approval, use, and ongoing monitoring. Organizations may have different model development guidelines across business units depending on the associated regulatory, legal, or business requirements.

5. Principles of Transparency

Explainability

14. Individuals should have access, upon request, to a reasonable explanation for any material decisions made by an AI application, or with AI application output. A reasonable explanation should at minimum include what data attributes are used, and how those data attributes impact the AI application output.

5.1. Organizations should align with AI stakeholders on the level of explainability required for an AI application prior to its approval. For all material AI applications, the organization should be able to explain what data attributes are used, and how those data attributes directionally impact the AI application output. The explanation should be understandable for a typical individual affected by that specific AI application. The agreed-upon level of explainability may determine what type of AI is acceptable for use.

5.2. An AI application may be deemed highly complex because of intentional protection and concealment, the reality that the skills for writing and reading artificial intelligence code are highly specialized, or an inability of humans to understand the highly complex outputs of a model[6]. Nevertheless, organizations are responsible for providing an explanation of what data attributes are used, and how those data attributes impact the AI application output for all material AI applications in production. AI application complexity will not be deemed an appropriate rationale for an inability to provide a reasonable explanation.

5.3. Intentional protection of an AI application for intellectual property or other reasons may be beneficial to external AI application providers. Extra due diligence may be required for organizations using external AI applications to ensure they meet the explainability requirements of the Code.

5.4. Example: A marketing manager is trying to source an AI application to help his team optimize advertising activities. He has two options, one built by a third party and another built internally. He consults with both teams of AI designers to understand what data attributes are being used and how those data attributes impact the recommendation for advertising activities, but the third party doesn't want to share all the details because they say, "that information is proprietary". The marketing manager knows that according to the Code, his organization must have access to that information and decides to proceed with the internal AI application, which is explainable.

[6] Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Data Privacy & Informed Consent

15. All AI applications should adhere to the data privacy laws, standards, and other related guidelines that govern the organization and industry.
16. Organizations should ensure individuals, internal and external, provide meaningful informed consent for the use of their data in circumstances where privacy is expected, per the applicable local laws.

5.5. Organizations are expected to adhere to all laws, standards and other related guidelines that govern the use of data in their industry. Entirely secure systems do not exist, so planning for a cyber attack and implementing safeguards is the best recommended defense.

5.6. Organizations should treat the data of all consumers, employees, and other individuals, with a similar level of prudence.

5.7. Organizations should ensure that all individuals have provided meaningful, informed consent for the use of their data in circumstances where privacy is expected per the applicable local laws. This includes ensuring individuals know what they are consenting to, ensuring the data is only used per the consent provided, showing consent has been provided, allowing individuals to withdraw consent and their data at any point, and ensuring a service or product's use is not conditional on providing consent unless doing so is deemed acceptable by the relevant governing bodies[7]. This applies to all data collected by the organization from this day forward, as data that has already been collected must often be retained for documentation purposes.

5.8. Example: A talent manager (AI owner) works with a data scientist (AI designer) to develop an automated resume reading AI application that evaluates potential candidates on a set of skill-based attributes. The data scientist and talent manager ask three members of the current HR team to review 100 successful and 100 unsuccessful historical resumes and rank them on the skill-based attributes. The talent manager must discuss the benefits and risks to the HR team members of providing their data, ensuring they understand, and ultimately provide informed consent. For example, a benefit could be that the AI application will significantly reduce the time spent reviewing resumes, but it may also create a risk of uncovering trends or behaviours in the HR employee's resume review techniques that have ethical implications.

[7] The European Parliament and The Council of The European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. *Official Journal of the European Union*, *59*(May), 1–88. https://doi.org/L:2016:119:TOC

Announcing the Use of AI Applications

17. Organizations should proactively announce the use of AI applications to individuals when they are interacting directly with an AI application, and when a material decision is made by an AI application or with AI application output.

5.9. Individuals may not know they are interacting with an AI application, especially if that interaction has historically occurred with another human. To maintain trust, organizations should ensure the individuals understand they are interacting with an AI application. When the AI application interacts directly with an individual, an announcement should occur each time there is a new interaction (i.e. each new chat-bot discussion). When the AI application doesn't interact directly with an individual, organizations may instead choose to announce the use of AI applications through more general channels (i.e. annual report, website).

5.10. Example: Many organizations allow customers to inquire about their services using virtual assistant AI applications such as Amazon's Alexa or Apple's Siri. In these instances, a customer is actively engaging with an AI application via their own virtual assistant AI application and are therefore aware they are communicating with an AI application. The same virtual assistant technology could soon be used by organizations to communicate with their own customers, and organizations will need to determine how they proactively announce that the customer is being contacted by an AI application, as the AI is replacing a historically human-to-human interaction.