< >
Montréal Declaration
Responsible AI_
</ >

MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

2018

This document is part of the 2018 **MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE.**
You can find the complete report HERE.

TABLE OF CONTENTS

READING THE DECLARATION

A DECLARATION, FOR WHAT PURPOSE?

The Montréal Declaration for responsible AI development has three main objectives:

1. Develop an ethical framework for the development and deployment of AI;
2. Guide the digital transition so everyone benefits from this technological revolution;
3. Open a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development.

A DECLARATION OF WHAT?

PRINCIPLES

The Declaration's first objective consists of identifying the ethical principles and values that promote the fundamental interests of people and groups. These principles applied to the digital and artificial intelligence field remain general and abstract. To read them correctly, it is important to keep the following points in mind:

- Although they are presented as a list, there is no hierarchy. The last principle is not less important than the first. However, it is possible, depending on the circumstances, to lend more weight to one principle than another, or to consider one principle more relevant than another.
- Although they are diverse, they must be interpreted consistently to prevent any conflict that could prevent them from being applied. As a general rule, the limits of one principle's application are defined by another principle's field of application.
- Although they reflect the moral and political culture of the society in which they were developed, they provide the basis for an intercultural and international dialogue.
- Although they can be interpreted in different ways, they cannot be interpreted in just any way. It is imperative that the interpretation be coherent.
- Although these are ethical principles, they can be translated into political language and interpreted in legal fashion.

Recommendations were made based on these principles to establish guidelines for the digital transition within the Declaration's ethical framework. It aims at covering a few key cross-sectorial themes to reflect on the transition towards a society in which AI helps promote the common good: algorithmic governance, digital literacy, digital inclusion of diversity and ecological sustainability.

A DECLARATION FOR WHOM?

The Montréal Declaration is addressed to any person, organization and company that wishes to take part in the responsible development of artificial intelligence, whether it's to contribute scientifically or technologically, to develop social projects, to elaborate rules (regulations, codes) that apply to it, to be able to contest bad or unwise approaches, or to be able to alert public opinion when necessary.

It is also addressed to political representatives, whether elected or named, whose citizens expect them to take stock of developing social changes, quickly establish a framework allowing a digital transition that serves the greater good, and anticipate the serious risks presented by AI development.

A DECLARATION ACCORDING TO WHAT METHOD?

The Declaration was born from an inclusive deliberation process that initiates a dialogue between citizens, experts, public officials, industry stakeholders, civil organizations and professional associations. The advantages of this approach are threefold:

1. Collectively mediate AI's social and ethical controversies;
2. Improve the quality of reflection on responsible AI;
3. Strengthen the legitimacy of the proposals for responsible AI.

The elaboration of principles and recommendations is a co-construction work that involved a variety of participants in public spaces, in the boardrooms of professional organizations, around international expert round tables, in research offices, classrooms or online, always with the same rigor.

AFTER THE DECLARATION?

Because the Declaration concerns a technology which has been steadily progressing since the 1950s, and whose pace of major innovations increases in exponential fashion, it is essential to perceive the Declaration as an open guidance document, to be revised and adapted according to the evolution of knowledge and techniques, as well as user feedback on AI use in society. At the end of the Declaration's elaboration process, we have reached the starting point for an open and inclusive conversation surrounding the future of humanity being served by artificial intelligence technologies.

PREAMBLE

For the first time in human history, it is possible to create autonomous systems capable of performing complex tasks of which natural intelligence alone was thought capable: processing large quantities of information, calculating and predicting, learning and adapting responses to changing situations, and recognizing and classifying objects. Given the immaterial nature of these tasks, and by analogy with human intelligence, we designate these wide-ranging systems under the general name of artificial intelligence. Artificial intelligence constitutes a major form of scientific and technological progress, which can generate considerable social benefits by improving living conditions and health, facilitating justice, creating wealth, bolstering public safety, and mitigating the impact of human activities on the environment and the climate. Intelligent machines are not limited to performing better calculations than human beings; they can also interact with sentient beings, keep them company and take care of them.

However, the development of artificial intelligence does pose major ethical challenges and social risks. Indeed, intelligent machines can restrict the choices of individuals and groups, lower living standards, disrupt the organization of labor and the job market, influence politics, clash with fundamental rights, exacerbate social and economic inequalities, and affect ecosystems, the climate and the environment. Although scientific progress, and living in a society, always carry a risk, it is up to the citizens to determine the moral and political ends that give meaning to the risks encountered in an uncertain world.

The lower the risks of its deployment, the greater the benefits of artificial intelligence will be. The first danger of artificial intelligence development consists in giving the illusion that we can master the future through calculations. Reducing society to a series of numbers and ruling it through algorithmic procedures is an old pipe dream that still drives human ambitions. But when it comes to human affairs, tomorrow rarely resembles today, and numbers cannot determine what has moral value, nor what is socially desirable.

The principles of the current declaration are like points on a moral compass that will help guide the development of artificial intelligence toward morally and socially desirable ends. They also offer an ethical framework that promotes internationally recognized human rights in the fields affected by the rollout of artificial intelligence. Taken as a whole, the principles articulated lay the foundation for cultivating social trust toward artificially intelligent systems.

The principles of the current declaration rest on the common belief that human beings seek to grow as social beings endowed with sensations, thoughts and feelings, and strive to fulfill their potential by freely exercising their emotional, moral and intellectual capacities. It is incumbent on the various public and private stakeholders and policymakers at the local, national and international level to ensure that the development and deployment of artificial intelligence are compatible with the protection of fundamental human capacities and goals, and contribute toward their fuller realization. With this goal in mind, one must interpret the proposed principles in a coherent manner, while taking into account the specific social, cultural, political and legal contexts of their application.

1

WELL-BEING PRINCIPLE

The development and use of artificial intelligence systems (AIS) must permit the growth of the well-being of all sentient beings.

1. AIS must help individuals improve their living conditions, their health, and their working conditions.

2. AIS must allow individuals to pursue their preferences, so long as they do not cause harm to other sentient beings.

3. AIS must allow people to exercise their mental and physical capacities.

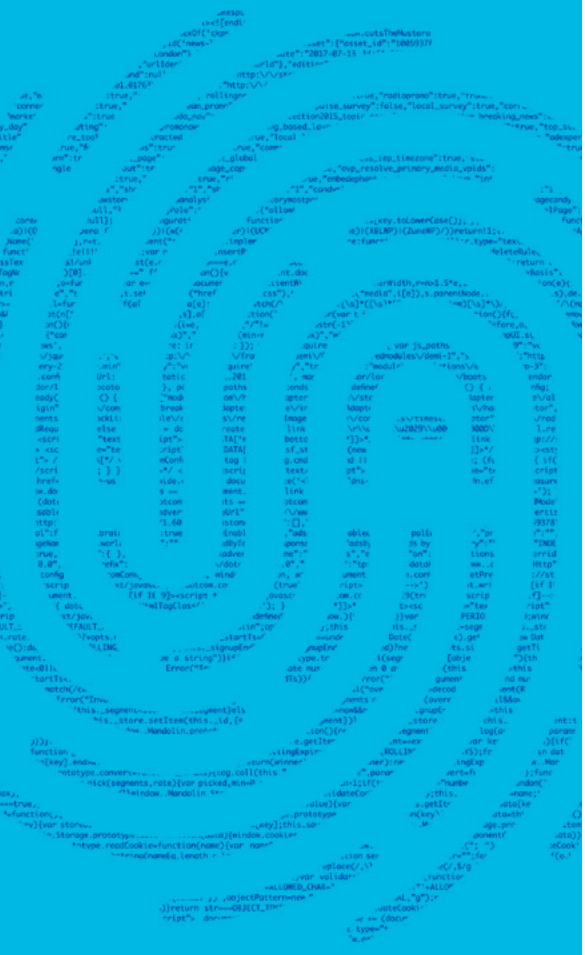
4. AIS must not become a source of ill-being, unless it allows us to achieve a superior well-being than what one could attain otherwise.

5. AIS use should not contribute to increasing stress, anxiety, or a sense of being harassed by one's digital environment.

2

RESPECT FOR AUTONOMY PRINCIPLE

AIS must be developed and used while respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.



1. AIS must allow individuals to fulfill their own moral objectives and their conception of a life worth living.

2. AIS must not be developed or used to impose a particular lifestyle on individuals, whether directly or indirectly, by implementing oppressive surveillance and evaluation or incentive mechanisms.

3. Public institutions must not use AIS to promote or discredit a particular conception of the good life.

4. It is crucial to empower citizens regarding digital technologies by ensuring access to the relevant forms of knowledge, promoting the learning of fundamental skills (digital and media literacy), and fostering the development of critical thinking.

5. AIS must not be developed to spread untrustworthy information, lies, or propaganda, and should be designed with a view to containing their dissemination.

6. The development of AIS must avoid creating dependencies through attention-capturing techniques or the imitation of human characteristics (appearance, voice, etc.) in ways that could cause confusion between AIS and humans.

3

PROTECTION OF PRIVACY AND INTIMACY PRINCIPLE

Privacy and intimacy must be protected from AIS intrusion and data acquisition and archiving systems (DAAS).

1. Personal spaces in which people are not subjected to surveillance or digital evaluation must be protected from the intrusion of AIS and data acquisition and archiving systems (DAAS).
2. The intimacy of thoughts and emotions must be strictly protected from AIS and DAAS uses capable of causing harm, especially uses that impose moral judgments on people or their lifestyle choices.
3. People must always have the right to digital disconnection in their private lives, and AIS should explicitly offer the option to disconnect at regular intervals, without encouraging people to stay connected.
4. People must have extensive control over information regarding their preferences. AIS must not create individual preference profiles to influence the behavior of the individuals without their free and informed consent.
5. DAAS must guarantee data confidentiality and personal profile anonymity.
6. Every person must be able to exercise extensive control over their personal data, especially when it comes to its collection, use, and dissemination. Access to AIS and digital services by individuals must not be made conditional on their abandoning control or ownership of their personal data.
7. Individuals should be free to donate their personal data to research organizations in order to contribute to the advancement of knowledge.
8. The integrity of one's personal identity must be guaranteed. AIS must not be used to imitate or alter a person's appearance, voice, or other individual characteristics in order to damage one's reputation or manipulate other people.

4

SOLIDARITY PRINCIPLE

The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.

1. AIS must not threaten the preservation of fulfilling moral and emotional human relationships, and should be developed with the goal of fostering these relationships and reducing people's vulnerability and isolation.

2. AIS must be developed with the goal of collaborating with humans on complex tasks and should foster collaborative work between humans.

3. AIS should not be implemented to replace people in duties that require quality human relationships, but should be developed to facilitate these relationships.

4. Health care systems that use AIS must take into consideration the importance of a patient's relationships with family and health care staff.

5. AIS development should not encourage cruel behavior toward robots designed to resemble human beings or non-human animals in appearance or behavior.

6. AIS should help improve risk management and foster conditions for a society with a more equitable and mutual distribution of individual and collective risks.

5

DEMOCRATIC PARTICIPATION PRINCIPLE

AIS must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control.

1. AIS processes that make decisions affecting a person's life, quality of life, or reputation must be intelligible to their creators.
2. The decisions made by AIS affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use. Justification consists in making transparent the most important factors and parameters shaping the decision, and should take the same form as the justification we would demand of a human making the same kind of decision.
3. The code for algorithms, whether public or private, must always be accessible to the relevant public authorities and stakeholders for verification and control purposes.
4. The discovery of AIS operating errors, unexpected or undesirable effects, security breaches, and data leaks must imperatively be reported to the relevant public authorities, stakeholders, and those affected by the situation.
5. In accordance with the transparency requirement for public decisions, the code for decision-making algorithms used by public authorities must be accessible to all, with the exception of algorithms that present a high risk of serious danger if misused.
6. For public AIS that has a significant impact on the life of citizens, citizens should have the opportunity and skills to deliberate on the social parameters of these AIS, their objectives, and the limits of their use.
7. We must at all times be able to verify that AIS are doing what they were programed for and what they are used for.
8. Any person using a service should know if a decision concerning them or affecting them was made by an AIS.
9. Any user of a service employing chatbots should be able to easily identify whether they are interacting with an AIS or a real person.
10. Artificial intelligence research should remain open and accessible to all.

6

EQUITY PRINCIPLE

The development and use of AIS must contribute to the creation of a just and equitable society.

1. AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on — among other things — social, sexual, ethnic, cultural, or religious differences.

2. AIS development must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge.

3. AIS development must produce social and economic benefits for all by reducing social inequalities and vulnerabilities.

4. Industrial AIS development must be compatible with acceptable working conditions at every step of their life cycle, from natural resources extraction to recycling, and including data processing.

5. The digital activity of users of AIS and digital services should be recognized as labor that contributes to the functioning of algorithms and creates value.

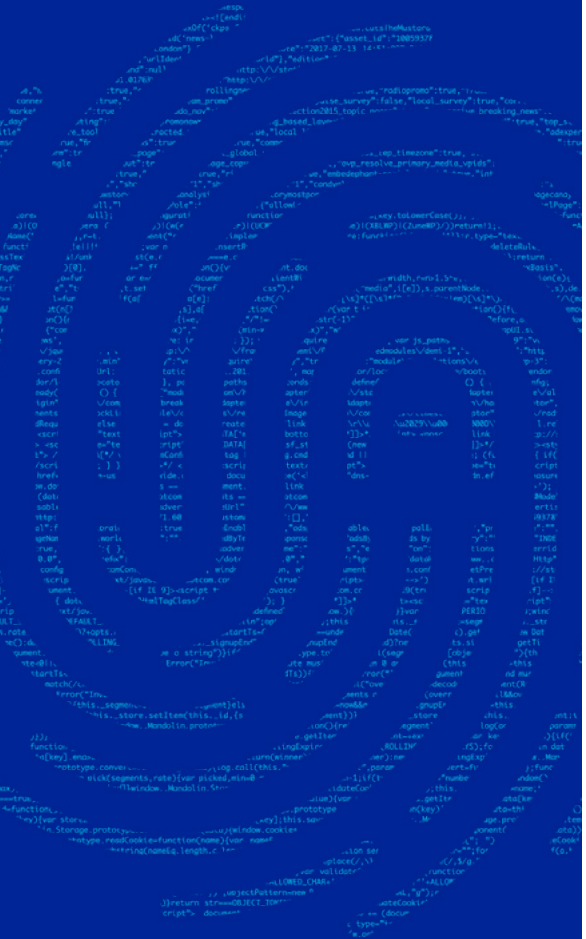
6. Access to fundamental resources, knowledge and digital tools must be guaranteed for all.

7. We should support the development of commons algorithms — and of open data needed to train them — and expand their use, as a socially equitable objective.

7

DIVERSITY INCLUSION PRINCIPLE

The development and use of AIS must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences.



1. AIS development and use must not lead to the homogenization of society through the standardization of behavior and opinions.

2. From the moment algorithms are conceived, AIS development and deployment must take into consideration the multitude of expressions of social and cultural diversity present in the society.

3. AI development environments, whether in research or industry, must be inclusive and reflect the diversity of the individuals and groups of the society.

4. AIS must avoid using acquired data to lock individuals into a user profile, fix their personal identity, or confine them to a filtering bubble, which would restrict and confine their possibilities for personal development — especially in fields such as education, justice, or business.

5. AIS must not be developed or used with the aim of limiting the free expression of ideas or the opportunity to hear diverse opinions, both being essential conditions of a democratic society.

6. For each service category, the AIS offering must be diversified to prevent de facto monopolies from forming and undermining individual freedoms.

8

PRUDENCE PRINCIPLE

Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AIS use and by taking the appropriate measures to avoid them.

1. It is necessary to develop mechanisms that consider the potential for the double use — beneficial and harmful — of AI research and AIS development (whether public or private) in order to limit harmful uses.

2. When the misuse of an AIS endangers public health or safety and has a high probability of occurrence, it is prudent to restrict open access and public dissemination to its algorithm.

3. Before being placed on the market and whether they are offered for charge or for free, AIS must meet strict reliability, security, and integrity requirements and be subjected to tests that do not put people's lives in danger, harm their quality of life, or negatively impact their reputation or psychological integrity. These tests must be open to the relevant public authorities and stakeholders.

4. The development of AIS must preempt the risks of user data misuse and protect the integrity and confidentiality of personal data.

5. The errors and flaws discovered in AIS and SAAD should be publicly shared, on a global scale, by public institutions and businesses in sectors that pose a significant danger to personal integrity and social organization.

9

RESPONSIBILITY PRINCIPLE

The development and use of AIS must not contribute to lessening the responsibility of human beings when decisions must be made.

1. Only human beings can be held responsible for decisions stemming from recommendations made by AIS, and the actions that proceed therefrom.

2. In all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decision must be taken by a human being and that decision should be free and informed.

3. The decision to kill must always be made by human beings, and responsibility for this decision must not be transferred to an AIS.

4. People who authorize AIS to commit a crime or an offense, or demonstrate negligence by allowing AIS to commit them, are responsible for this crime or offense.

5. When damage or harm has been inflicted by an AIS, and the AIS is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use.

10

SUSTAINABLE DEVELOPMENT PRINCIPLE

The development and use of AIS must be carried out so as to ensure a strong environmental sustainability of the planet.

1. AIS hardware, its digital infrastructure and the relevant objects on which it relies such as data centers, must aim for the greatest energy efficiency and to mitigate greenhouse gas emissions over its entire life cycle.

2. AIS hardware, its digital infrastructure and the relevant objects on which it relies, must aim to generate the least amount of electric and electronic waste and to provide for maintenance, repair, and recycling procedures according to the principles of circular economy.

3. AIS hardware, its digital infrastructure and the relevant objects on which it relies, must minimize our impact on ecosystems and biodiversity at every stage of its life cycle, notably with respect to the extraction of resources and the ultimate disposal of the equipment when it has reached the end of its useful life.

4. Public and private actors must support the environmentally responsible development of AIS in order to combat the waste of natural resources and produced goods, build sustainable supply chains and trade, and reduce global pollution.

GLOSSARY

Algorithm
An algorithm is a method of problem solving through a finite and non-ambiguous series of operations. More specifically, in an artificial intelligence context, it is the series of operations applied to input data to achieve the desired result.

Artificial intelligence (AI)
Artificial intelligence (AI) refers to the series of techniques which allow a machine to simulate human learning, namely to learn, predict, make decisions and perceive its surroundings. In the case of a computing system, artificial intelligence is applied to digital data.

Artificial intelligence system (AIS)
An AIS is any computing system using artificial intelligence algorithms, whether it's software, a connected object or a robot.

Chatbot
A chatbot is an AI system that can converse with its user in a natural language.

Data Acquisition and Archiving System (DAAS)
DAAS refers to any computing system that can collect and record data. This data is eventually used to train AI systems or as decision-making parameters.

Decision Justifiability
An AIS's decision is justified when there exist non-trivial reasons that motivate this decision, and that these reasons can be communicated in natural language.

Deep Learning
Deep learning is the branch of machine learning that uses artificial neuron networks on many levels. It is the technology behind the latest AI breakthroughs.

Digital Commons
Digital commons are the applications or data produced by a community. Unlike material goods, they are easily shareable and do not deteriorate when used. Therefore, unlike proprietary software, open source software—which is often the result of a collaboration between programmers—are considered digital commons since their source code is open and accessible to all.

Digital Disconnection
Digital disconnection refers to an individual's temporary or permanent ceasing of online activity.

Digital Literacy
An individual's digital literacy refers to their ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital tools and networked technologies to participate in economic and social life.

Filter Bubble
The filter bubble (or filtering bubble) expression refers to the "filtered" information which reaches an individual on the Internet. Various services such as social networks or search engines offer personalized results for their users. This can have the effect of isolating individuals (inside "bubbles") since they no longer have access to common information.

GAN
Acronym for Generative Adversarial Network. In a GAN, two antagonist networks are placed in competition to generate an image. They can for example be used to create an image, a recording or a video that appears practically real to a human being.

Intelligibility
An AIS is intelligible when a human being with the necessary knowledge can understand its operations, meaning its mathematical model and the processes that determine it.

Machine Learning
Machine learning is the branch of artificial intelligence that consists of programing an algorithm so that it can learn by itself.

The various techniques can be classified into three major types of machine learning:

> In supervised learning, the artificial intelligence system (AIS) learns to predict a value from entered data. This requires annotated entry-value couples during training. For example, a system can learn to recognize an object featured in a picture.

> In unsupervised learning, AIS learns to find similarities among data that hasn't been annotated, for example in order to divide them into various homogeneous partitions. A system can thereby recognize communities of social media users.

> Through reinforcement learning, AIS learns to act on its environment in order to maximize the reward it receives during training. This is the technique through which AIS was able to beat humans in the game of Go or the videogame Dota2.

Online Activity
Online activity refers to all activities performed by an individual in a digital environment, whether those activities are done on a computer, a telephone or any other connected object.

Open Data
Open data is digital data that users can access freely. For example, this is the case for most published AI research results.

Path Dependency
Social mechanism through which technological, organizational or institutional decisions, once deemed rational but now subpar, still continue to influence decision-making. A mechanism maintained because of cognitive bias or because change would require too much money or effort. Such is the case for urban road infrastructure when it leads to traffic optimization programs, rather than considering a change to organize transportation with very low carbon emissions. This mechanism must be known when using AI for special projects, as training data in supervised learning can sometimes reinforce old organizational paradigms that are now contested.

Personal Data
Personal data are those that help directly or indirectly identify an individual.

Rebound Effect
The rebound effect is the mechanism through which greater energy efficiency or better environmental performance of goods, equipment and services leads to an increase in use that is more than proportional. For example, screen size increases, the number of electronic devices in a household goes up, and greater distances are traveled by car or plane. The global result is greater pressure on resources and the environment.

Reliability
An AIS is reliable when it performs the task it was designed for, in expected fashion. Reliability is the probability of success that ranges between 51% and 100%, meaning strictly superior to chance. The more a system is reliable, the more its behavior is predictable.

Strong Environmental Sustainability

The notion of strong environmental sustainability goes back to the idea that in order to be sustainable, the rate of natural resource consumption and polluting emissions must be compatible with planetary environmental limits, the rate of resources and ecosystem renewal, and climate stability.

Unlike weak sustainability, which requires less effort, strong sustainability does not allow the substitution of the loss of natural resources with artificial capital.

Sustainable Development

Sustainable development refers to the development of human society that is compatible with the capacity of natural systems to offer the necessary resources and services to this society. It is economic and social development that fulfills current needs without compromising the existence of future generations.

Training

Training is the machine learning process through which AIS build a model from data. The performance of AIS depends on the quality of the model, which itself depends on the quantity and quality of data used during training.

CREDITS

The writing of the Montréal Declaration for the responsible development of artificial intelligence is the result of the work of a multidisciplinary and inter-university scientific team that draws on a citizen consultation process and a dialogue with experts and stakeholders of AI development.

Christophe Abrassart, Associate Professor in the School of Design and Co-director of Lab Ville Prospective of the Faculty of Planning of the Université de Montréal, member of *Centre de recherche en éthique* (CRÉ)

Yoshua Bengio, Full Professor of the Department of Computer Science and Operations Research, UdeM, Scientific Director of MILA and IVADO

Guillaume Chicoisne, Scientific Programs Director, IVADO

Nathalie de Marcellis-Warin, Full Professor, Polytechnique Montréal, President and Chief Executive officer, Center for Interuniversity Research and Analysis of Organizations (CIRANO)

Marc-Antoine Dilhac, Associate Professor, Department of Philosophy, Université de Montréal, Chair of the Ethics and Politics Group, *Centre de recherche en éthique* (CRÉ), Canada Research Chair in Public Ethics and Political Theory, Director of the Institut Philosophie Citoyenneté Jeunesse

Sébastien Gambs, Professor of Computer Science of Université du Québec à Montréal, Canada Research Chair in Privacy-Preserving and Ethical Analysis of Big Data

Vincent Gautrais, Full Professor, Faculty of Law, Université de Montréal; Director of the *Centre de recherche en droit public* (CRDP); Chair of the L.R. Wilson Chair in Information Technology and E-Commerce Law

Martin Gibert, Ethics Counsellor at IVADO and researcher in *Centre de recherche en éthique* (CRÉ)

Lyse Langlois, Full Professor and Vice-Dean of the Faculty of Social Science; Director of the *Institut d'éthique appliquée* (IDÉA); Researcher Interuniversity Research Center on Globalization and Work (CRIMT)

François Laviolette, Full Professor, Department of Computer Science and Software Engineering, Université Laval; Director of the Centre de recherche en données massives (CRDM)

Pascale Lehoux, Full Professor at the *École de santé publique*, Université de Montréal (ESPUM); Chair on Responsible Innovation in Health

Jocelyn Maclure, Full Professor, Faculty of Philosophy, Université Laval, and President of the Quebec Ethics in Science and Technology Commission (CEST)

Marie Martel, Professor in *École de bibliothéconomie et des sciences de l'information*, Université de Montréal

Joëlle Pineau, Associate Professor, School of Computer Science, McGill University; Director of Facebook AI Lab in Montréal; Co-director of the Reasoning and Learning Lab

Peter Railton, Gregory S. Kavka Distinguished University Professor; John Stephenson Perrin Professor; Arthur F. Thurnau Professor, Department of Philosophy, University of Michigan, Fellow of the American Academy of Arts & Sciences

Catherine Régis, Associate professor, Faculty of Law, Université de Montréal; Canada Research Chair in Collaborative Culture in Health Law and Policy; Regular researcher, *Centre de recherche en droit public* (CRDP)

Christine Tappolet, Full Professor, Department of Philosophy, UdeM, Director of Centre de recherche en éthique (CRÉ)

Nathalie Voarino, PhD Candidate in Bioethics of Université de Montréal

OUR PARTNERS

Université de Montréal

CRE
CENTRE DE RECHERCHE EN ETHIQUE

LAB VILLE PROSPECTIVE

IVADO
INSTITUT DE VALORISATION DES DONNÉES

SAT

Mila

ICRA
Programme
IA et
société

Université de Montréal | designNsociété

polEtHics
CHAIRE DE RECHERCHE DU CANADA
ÉTHIQUE PUBLIQUE ET THÉORIE POLITIQUE

CIRANO
Allier savoir et décision
Centre interuniversitaire de recherche en analyse des organisations

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

UNIVERSITÉ LAVAL
Institut d'éthique appliquée

MUSÉE DE LA CIVILISATION
Québec

CENTRE D'ÉTUDES ET DE RECHERCHES INTERNATIONALES
CÉRIUM
Université de Montréal

crdm_ul
CENTRE DE RECHERCHE EN DONNÉES MASSIVES DE L'UNIVERSITÉ LAVAL

Canada

Centre Culturel Canadien Paris
Canadian Cultural Centre Paris

MEC
Maison des étudiants canadiens

< >

Montréal Declaration Responsible AI_

</ >

montrealdeclaration-responsibleai.com