< >

Montréal Declaration
Responsible AI_

</ >

2018 REPORT

MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

TABLE OF CONTENTS

MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE 3

PART 1
CO-CONSTRUCTION APPROACH AND METHODOLOGY 22

PART 2
2018 OVERVIEW OF INTERNATIONAL RECOMMENDATIONS FOR AI ETHICS 78

PART 3
SUMMARY REPORT OF RECOMMENDATIONS FROM THE WINTER CO-CONSTRUCTION WORKSHOPS 98

PART 4
FALL 2018 CO-CONSTRUCTION: KEY ACTIVITIES 150

PART 5
SUMMARY REPORT OF ONLINE SURVEYS AND PROPOSALS RECEIVED FOR THE MONTRÉAL RESPONSIBLE AI DECLARATION 212

PART 6
PRIORITY PROJECTS AND THEIR RECOMMENDATIONS FOR RESPONSIBLE AI DEVELOPMENT 249

CREDITS I

PARTNERS II

< >
Montréal Declaration
Responsible AI_
</ >

MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

2018

TABLE OF CONTENTS

READING THE DECLARATION		5
PREAMBLE		7
PRINCIPLES		
1.	WELL-BEING PRINCIPLE	8
2.	RESPECT FOR AUTONOMY PRINCIPLE	9
3.	PROTECTION OF PRIVACY AND INTIMACY	10
4.	SOLIDARITY PRINCIPLE	11
5.	DEMOCRATIC PARTICIPATION PRINCIPLE	12
6.	EQUITY PRINCIPLE	13
7.	DIVERSITY INCLUSION PRINCIPLE	14
8.	CAUTION PRINCIPLE	15
9.	RESPONSIBILITY PRINCIPLE	16
10.	SUSTAINABLE DEVELOPMENT PRINCIPLE	17
GLOSSARY		18
CREDITS		I

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 in 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 no 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.

From those principles were elaborated some recommendations the purpose of which is to suggest guidelines to accomplish 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 rigour.

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 towards 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 towards 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.

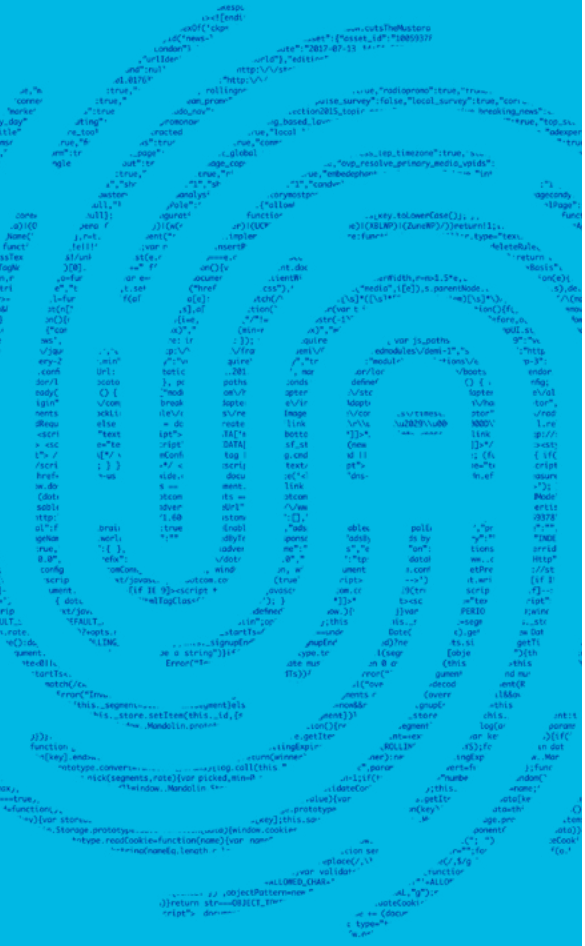
4. AIS must not become a source of ill-being, unless it allows us to achieve a superior well-being than what one could attain otherwise.

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 have 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 programmed 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.

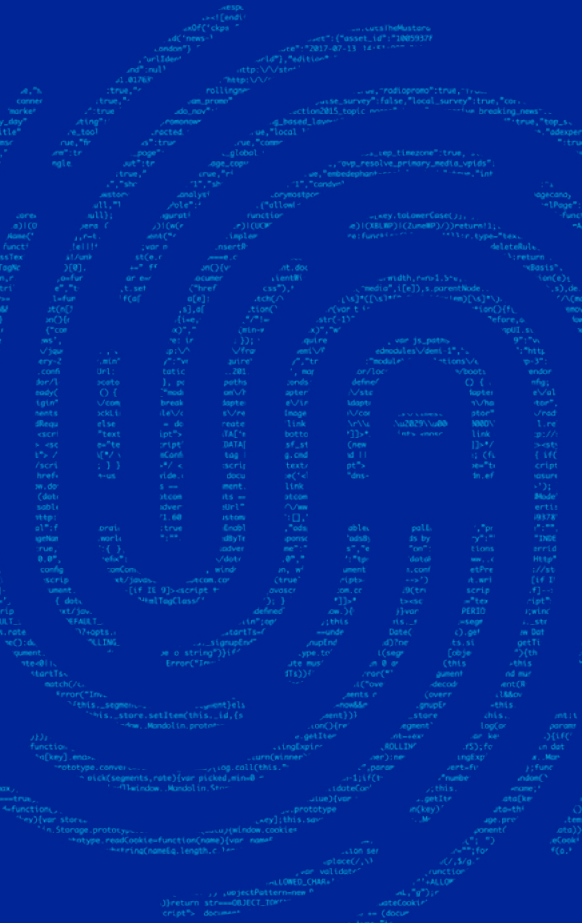
6. Access to fundamental resources, knowledge and digital tools must be guaranteed for all.

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 behaviours 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 of which 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 lessen 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 offence, or demonstrate negligence by allowing AIS to commit them, are responsible for this crime or offence.

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 centres, 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 disposition 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 programming 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 on a picture.

> In unsupervised learning, AIS learns to find similarities amongst 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 were 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 of 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 travelled 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 behaviour is predictable.

Strong Environmental Sustainability

The notion of strong environmental sustainability goes back to the idea that in order to be sustainable, the rhythm of natural resource consumption and polluting emissions must be compatible with planetary environmental limits, the rhythm 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 Reasearch 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); Chairholder of the L.R. Wilson Chair in Information Technology and E-Commerce Law

Martin Gibert, Ethics Counsellor at IVADO et 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 Centre on Globalization and Work (CRIMT)

François Laviolette, Full Professor, au 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 School of Public Health of University of Montreal (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; Chairholder, Canada Research Chair in Collaaborative 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

< >

Montréal Declaration
Responsible AI_

</ >

PART 1

CO-CONSTRUCTION APPROACH AND METHODOLOGY

TABLE OF CONTENTS

SUMMARY		25
1.	**INTRODUCTION**	**28**
2.	**Why have a *Montréal Declaration* for Responsible AI?**	**31**
2.1	The intellectual origins of this project	32
2.2	Forum on the Socially Responsible Development of Artificial Intelligence	34
2.3	Towards the *Montréal Declaration* for Responsible AI Development	35
2.4	Montréal and the international context	36
3.	**THE ETHICAL AND SOCIAL ISSUES OF AI**	**38**
3.1	What is AI?	38
3.2	AI in everyday life and philosophical questioning	39
3.3	The Ethical issues of AI	41
3.4	AI Ethics and the *Montréal Declaration*	42
4.	**THE CO-CONSTRUCTION APPROACH**	**44**
4.1	The principles of the co-construction approach	44
	4.1.1 The Principles of Good Citizen Involvement	44
	4.1.2 Experts and Citizens	46
4.2	The Co-construction workshop methodology	47
4.3	Uniqueness of the co-construction approach	48
4.4	World cafés outside libraries	50
4.5	Portrait of participants	50

CONTRIBUTORS

MARC-ANTOINE DILHAC, Scientific Co-director of the Declaration, Full Professor, Department of Philosophy, UdeM, Chair of the Ethics and Politics Group, Centre de recherche en éthique (CRÉ), Canada Research Chair in Public Ethics and Political Theory

CHRISTOPHE ABRASSART, Scientific Co-director of the Declaration, 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)

TABLE OF CONTENTS

5. WORKSHOP DELIBERATIONS: EXAMPLES FROM SMART CITIES AND THE WORKPLACE 53

5.1 The deliberation process 53

5.1.1 Smart city sector: self-driving cars (SDC) and sharing the road equitably 54

5.1.2 Workplace sector: Socially responsible restructuring? 58

6. PARTICIPANTS IN THE CO-CONSTRUCTION AND WORKING GROUPS 63

ANNEXES 68

Annex 1
Co-construction workshops: Detailed description of how they work 68

Annex 2
Foresight scenarios: winter co-construction workshops 70

TABLE AND FIGURES

Figure 1: The Values of the Declaration (preliminary version) 31

Figure 2: The Values of the *Montréal Declaration* for a Responsible Development of Artificial Intelligence 32

Figure 3: The co-construction approach 35

Figure 4: Strategic forecasting: a three-step process 49

Figure 5: Proportion of men and women involved in the co-construction workshops 50

Figure 6: Participants in the co-construction workshops per age group 50

Figure 7: Distribution of participants in world cafés and co-construction days by education level reached 51

Figure 8: Distribution of participants in world cafés and co-construction days by field of activity 52

Table 1: Smart City, First deliberative moment: formulating ethical issues in 2025 55

Table 2: Smart City, Second deliberative moment: AI framework recommendations for 2018-2020 56

Table 3: Workplace, First deliberative moment: formulating ethical issues in 2025 60

Table 4: Workplace, Second deliberative moment: recommendations for an AI framework in 2018-2020 61

Table 5: Typical procedure for world cafés 68

Table 6: Typical procedure for co-construction days 69

Table 7: Scenario summaries 71

Table 8: Elements of five scenarios 73

SUMMARY

On November 3, 2017, the Université de Montréal launched the co-construction process for the *Montréal Declaration* for a Responsible Development of Artificial Intelligence (*Montréal Declaration*). A year later, we present the results of these citizen deliberations. Dozens of events were organized to stimulate discussion on social issues that arise with artificial intelligence (AI), and 15 deliberation workshops were held over three months, involving over 500 citizens, experts and stakeholders from all backgrounds.

The *Montréal Declaration* is a collective endeavour that aims to steer the development of AI to support the common good, and guide social change by making recommendations with a strong democratic legitimacy.

The selected citizen co-construction method is based on a preliminary declaration of general ethical principles structured around seven (7) fundamental values: well-being, autonomy, justice, privacy, knowledge, democracy and responsibility. Following the process, the Declaration was enriched and now presents 10 principles based on the following values: well-being, autonomy, intimacy and privacy, solidarity, democracy, equity, inclusion, caution, responsibility and environmental sustainability.

If one of the goals of the co-construction process is to fine-tune the ethical principles suggested in the preliminary version of the *Montréal Declaration*, an equally important goal consists of making recommendations to provide a framework for AI research, as well as its technological and industrial development.

First, what is AI?

Very briefly, AI consists of simulating certain learning processes of human intelligence, learning from them and replicating them. For example, identifying complex patterns among a large quantity of data, or reasoning in a probabilistic fashion to sort information into categories, predict quantitative data or aggregate data. These cognitive skills form the basis for other skills such as choosing among several possible actions to reach a goal, interpreting an image or a sound, predicting a behaviour, anticipating an event, diagnosing a pathology and more. Two elements are key to these AI feats: data and algorithms, a series of instructions that perform a complex action.

To discuss the ethical issues of AI in concrete terms, the **co-construction workshop method** is based on the preliminary version of the *Montréal Declaration*. Schematically, after deciding on the "why?" (which desirable ethical principles should be included in a declaration on the ethics of AI?), it then becomes a matter of envisioning, along with participants, how ethical issues in the fields of health, justice, smart cities, education and culture, workplace and public services could arise in upcoming years. Then, we think about how we could respond to these issues. For example, through measures such as sector certification, a new stakeholder/mediator, a form or standard, public policy or a research program.

Citizens and stakeholders took part in world cafés or entire co-construction days where they had the chance to debate prospective scenarios.

Other citizens chose to contribute to the reflection by filling out an online questionnaire or submitting a paper. The results of these specific initiatives will be discussed in the global report on the activities associated with the *Montréal Declaration*, which should be published in the fall of 2018

.

Co-construction workshop results — General trends

In general, participants recognized that the advent of AI also brought important potential benefits. Participants especially recognized the time savings that AI devices could bring to their fields of work. However, they also felt that caution should be exercised in AI development to prevent it being abused or used for harmful purposes.

The citizens highlighted the need to implement different mechanisms to ensure that quality, understandable, transparent and relevant information was communicated. They also discussed the difficulty of guaranteeing truly enlightened consent.

The majority of participants recognized the need to align private and public interests, prevent monopolies from emerging and limit the influence of corporations.

Participants also recommended introducing mechanisms that would come from and involve people independently trained in technological and ethical issues of digital transition and AI to promote diversity, include the most vulnerable and protect the plurality of lifestyles.

No matter what it was used for, most participants insisted that AI remain a tool, and that final decisions be made by human beings when fundamental issues are at stake.

Priorities according to the principles of the *Montréal Declaration*

The principle of responsibility was considered the most pressing issue, followed by respect of autonomy and protection of privacy. Well-being, knowledge and justice came next. It should be noted, however, that they are all closely linked.

The principle of autonomy, considered a priority by most participants, entails respecting and promoting individual autonomy when they risk being controlled by technology and becoming dependent on tools. It also raises the issue of freedom of choice being two-sided: being able to make your own choice when faced with an AI-guided decision as well as the choice to not use these tools without risking social exclusion.

Participants also felt the principle of well-being was important. It was implicit at every roundtable, illustrating a collective desire to move towards a just and equitable society that fosters the development of all individuals. Overall, experts and users in every field concurred that the principle of well-being also serves as a reminder to maintain authentic human and emotional relationships.

Issues that led to creating new principles, or deliberating and exploring new themes

The principle of justice was discussed as the basis for two types of issues, which could lead to two new principles: a **principle of diversity**, which seeks to avoid discrimination by identifying bias-free mechanisms and a **principle of equity or social justice**, which states that AI benefits be accessible to all, and that its development not contribute to growing economic and social inequalities, but rather help bridge the gap.

A principle of caution. Issues related to trust in the development of AI technologies were regularly raised. This issue of trust is also closely tied to the question of reliability of AI systems.

A principle of explainability or justifiability. This principle implies being able to understand an algorithmic decision and react to it. For this, citizens felt it was important that algorithmic procedures be explained so they could see and understand which criteria were considered in the decision.

A principle of environmental sustainability. The impact of AI development and use on the environment raises specific issues, namely how to guarantee the responsible and fair use of material and natural resources.

Mechanisms for a digital transition

All the co-construction roundtables agreed on **three (3) priority mechanisms** to ensure socially responsible AI development, regardless of the field:

1. Include legal provisions.
2. Provide everyone with training.
3. Identify key independent stakeholders for AI management.

Pursuing the deliberations

The *Montréal Declaration* project focussed its first year of co-construction on many key sectors: education, health, work, smart city, predictive policing, environment, democracy and media propaganda. It is clear that a year of co-construction cannot possibly cover all the ethical and social issues associated with AI. The *Montréal Declaration* is not only the result of a collective reflection process, it is the very process itself: beyond Year 1 of the *Montréal Declaration*, the collective consultation and reflection process continues, because technological evolution waits for no one.

We present public policy recommendations around priority action areas. To date, four priority areas have emerged: algorithmic governance, digital literacy, diversity and inclusion, and environmental transition.

1. INTRODUCTION

On November 3, 2017, the Université de Montréal, in collaboration with the Fonds de recherche du Québec, launched a co-construction process based on the **Montréal Declaration for a Responsible Development of Artificial Intelligence** (*Montréal Declaration*). We had no idea the level of interest this initiative would generate, nor of the size of the task that lay ahead. A year later, we present the results of citizen deliberations, which involved various groups from civil society, citizens, experts, professional bodies, industry stakeholders and policymakers. It was a resounding success: dozens of events were held to discuss social issues surrounding AI, and 15 deliberation workshops took place from February to October, involving over 500 citizens, experts and stakeholders of all professional backgrounds.

The report we are presenting must be taken as a summary of a democratic deliberation process to enlighten public policy decisions on artificial intelligence, an experience which can serve as a reference point for other deliberative forums. The work on what is called the *Montréal Declaration* was led by a multidisciplinary and inter-university team of researchers, mainly in Quebec but also across the world. Awareness of social issues around artificial intelligence is shared not only by this research community, but by society as a whole. We suggested a citizen co-construction process because we are convinced that everyone has a right to be heard about how our society should be. This approach is innovative in both content and form: first, because it introduces foresight methods of applied ethics, which consist of anticipating ethical controversies around future artificial intelligence technologies or social situations where these technologies are used in unprecedented ways. Following this, we carried out this consultation process on a vastly broader scope. The numbers mentioned above paint a clear picture. This process should continue beyond the public presentation of the *Montréal Declaration*, since it must remain open to review.

We solicited the public to draft the Declaration; in return, were asked the following questions by not only the public, but various stakeholders: What will the Declaration change? Who is writing it? Isn't this just a vain university endeavour? Aren't there already too many manifestos, professions of faith on the ethical values of artificial intelligence? Isn't developing artificial intelligence within a framework of ethical principles and recommendations a means of condoning it? Isn't that approving a technocratic vision of society? Why not devote our energy to criticizing this development? None of these questions are without merit, and because we are committed to fostering greater transparency around artificial intelligence, we are also committed to increased transparency around the process we established. Our hope is that this report will provide a few answers.

The ethics of artificial intelligence have been a hot topic in many countries over the last two years. Each stakeholder in its development, not to mention researchers, businesses, citizens and political representatives, recognize the urgency of establishing an ethical, political and legal framework to guide the research and use of artificial intelligence. There is no doubt that with the rise of artificial intelligence technologies we are at the dawn of a new industrial revolution. The impact of this revolution on the production of goods, delivery of services, organization of work and the workforce, or even on family and personal relationships are still unknown but will be major, and possibly disruptive in certain fields. The social changes triggered by artificial intelligence are, indeed, surprisingly swift and illicit varied reactions, from enthusiasm to disapproval and scepticism. We could simply ignore them and embark in speculation about whether or not what we call artificial intelligence exists, but we'd only be postponing the problem to a point when it would no longer be possible to influence its development.

A number of objections and fears were raised during this first co-construction process. Many workshop participants and observers from the Declaration project questioned the technocratic ideology that views technology as a way of rationally organizing all of society, thus reducing social issues to technical problems. Others questioned the ability and the will of public institutions to regulate lucrative technologies. These objections cannot be casually dismissed, because they are based on historical precedents that shook our confidence in technological innovations, all the more so in the people promoting them. It is also important that individuals who raise objections do not undermine efforts to positively influence the future of society, but support them by getting involved in the democratic deliberations that allow us to maintain control over the development of digital technologies and artificial intelligence. We can complain about the effects of these new technologies on social relationships, or criticize how social life is being reduced to a series of lifestyles, but this will not prevent technological innovation, nor will it influence it. Yet that is the entire purpose of the *Montréal Declaration*: guide the development of artificial intelligence in order to promote fundamental ethical and social interests and provide guidelines for protecting human rights.

To conclude, we are not presenting a theory of artificial intelligence in this report, nor are we defending sophisticated arguments to settle the unrelenting question on the use of the term "artificial intelligence": is it an appropriate term to refer to data processing, recognition and decision-making algorithms? Some contest the use of this term by arguing that artificial intelligence refers to very limited knowledge processes when compared with human intelligence, or even the behavioural intelligence of pigeons. That is undeniable. But in that case, then, we must also recognize that complexity of paramecia surpasses that of any algorithm, even a learning one. If we go down that path, we will merely encounter roadblocks to understanding intelligence as a whole. What is human intelligence? Is there one or many forms? Do we need to introduce and specify an "emotional" form of intelligence? And in that case, why refuse to introduce an "artificial" form of intelligence? The hundreds of thousands of pages that have been written to answer these questions still do not suffice.

However, a few statements can help clear up misunderstandings that are at the root of the controversy: First, we know that the way biological neural networks operate is vastly different from that of artificial neural networks; there is no mistaking the two. But that does not invalidate the use of the term "artificial intelligence". If that were the case, the term mechanical arm would have to be discarded as well, given that a biological arm operates very differently, and that bones, joints, tendons and muscles are not pieces of metal, pulleys, springs and ropes. In general, people often confuse intelligence and thought. Intelligence is a property of thought, it is not thought as a whole. Intelligence, therefore, is particular in that it reduces the complexity of the world in which intelligent beings evolve to allow them to better master their environment. We give ourselves rules to analyze, calculate, evaluate and make decisions about reality. A long philosophical tradition of highly intelligent thinkers have asserted this, from Socrates to Russell to Leibniz. In a certain way, intelligence models and reduces reality to better act on it, like a mechanical equation models

and reduces movement to better understand it. Consequently, given the above, intelligence, even human, is largely algorithmic: it analyzes data and makes calculations according to procedures. It then lends itself to "mechanization" and "incarnation" very well, in the literal sense of the term: digital calculation, meaning calculations made on and with the fingers according to very diverse techniques, is an incarnation of calculations; with different abacuses such as the Chinese abacus, the Pascaline[1] and the electronic calculator, we witness the mechanization of calculation.

Reflecting on the goals we wish to pursue is not strictly a matter of calculations. Building your personal and social life around certain worthwhile goals does not depend on an algorithmic function. Knowing if we must use nuclear weapons to kill the greatest number of people and weaken an enemy country cannot be solely determined by calculating the consequences; we must still define the good or the goods according to which the calculation of consequences has a moral sense. There's something tragic about avoiding reflection on moral consequences by seeking only a calculation of the means. Artificial intelligence cannot yet engage in this kind of reflection. In the world we know and can anticipate in the near and mid-term future, reflecting on the finality of social life and existence in general is still a product of human intelligence.

The *Montréal Declaration* rests entirely on this statement: it is up to human and collective intelligence to define the purposes of social life and thereby, guide the development of artificial intelligence so that it is socially responsible and morally acceptable.

[1] Mechanical calculator designed and presented by mathematician and philosopher Blaise Pascal in 1645.

2. WHY HAVE A MONTRÉAL DECLARATION FOR RESPONSIBLE AI?

The *Montréal Declaration* is a collective work that aims to accomplish three (3) goals:

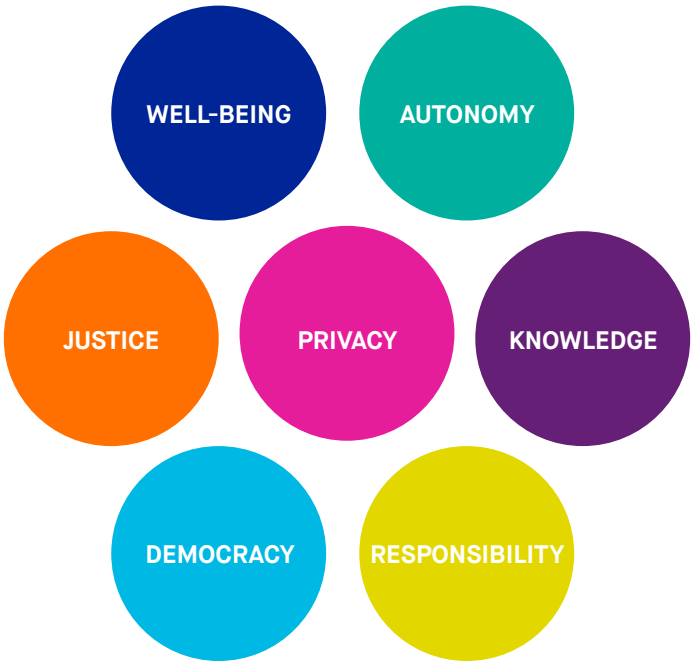
1. develop an ethical framework for ethical AI development and deployment
2. guide the digital transition so everyone benefits from this technological revolution
3. create a forum for national and international dialogue to successfully pursue inclusive and equitable AI development.

It becomes, therefore, a question of using AI development to ensure everyone's well-being, and guiding social change by developing recommendations founded in democratic legitimacy.

The Declaration is the outcome of an inclusive deliberation process that opens new dialogue between citizens, experts, public officials, industry stakeholders, civilian organizations and professional bodies.

The selected citizen co-construction method is based on a preliminary declaration of general ethical principles that is structured around fundamental values.

Figure 1: The Values of the Declaration (preliminary version)



Our relationship to these "values" is then broken down into standards we call principles. For example, if the value is well-being, our relationship to this value is that of maximization: we must increase the well-being of sentient beings. If the value is autonomy, our relationship is that of respect or protection: we must respect the autonomy of moral beings. The goal of the initial task of identifying these values and principles was to launch a citizen participation process that would then define the ethical principles of responsible AI development and recommendations to implement to ensure that AI promotes fundamental human interests.

At the end of this process, the values and principles were fine-tuned, allowing us to pinpoint things more precisely:

Figure 2: The Values of the Montréal Declaration for a Responsible Development of Artificial Intelligence

Moving from a preliminary to final version of the Declaration resulted from discussions that arose during the public consultation and co-construction workshops. The choice of values and principles rests on an understanding of fundamental social expectations as they were expressed, and is motivated by a desire to examine priority issues and find balance between the different values for the sake of coherency. Because there is no template formula to select principles (no algorithm yet exists for this task), it is the result of a complex adjustment process generally referred to as deliberation.

2.1

THE INTELLECTUAL ORIGINS OF THIS PROJECT

The revolution in artificial intelligence (AI), and more specifically deep learning, opens our perspectives to unimagined technological developments that will help improve decision-making, reduce certain risks and help the most vulnerable. This revolution is remarkable in many ways, although it also brings up questions that were first raised in the 18th century during the Industrial Revolution. It would be unwise to ignore the unique aspect of this revolution by hiding behind platitudes that leave us ill-prepared to face current challenges. Of course, human beings are gifted beings with great technical abilities—human history is itself a history of technical transformations of nature, and artificial intelligence extends this trend to automation—but upon closer inspection nothing resembles what is at stake today with the arrival of artificial intelligence technologies. The cognitive skills we believed unique to humans can now be performed by algorithms, machines that must be recognized, in a certain sense, as intelligent.

Although the social impact of these new technologies is wide-ranging, it is still somewhat unknown. It could prove disastrous if we do not take the time now to think about the ethical, political, judicial, sociological or psychological ramifications on the type of society and human relationships we want to promote or protect while still benefiting from information technologies and algorithm calculations.

Using algorithms to make technical or administrative decisions is nothing new. Although algorithms have been around since the Middle Ages[2], the rise of decision-making algorithms truly began in the 1950s, especially in the field of healthcare: emergency room triage in hospitals, detection of sudden infant death syndrome risks, prediction of heart attacks[3]. All these algorithm techniques—"procedures"—already raise a certain number of ethical and social issues: those

2 Algorithmic procedures have been known since Antiquity in fact, but contrary to what the "th" in algorithm may lead to believe, the word does not come from Ancient Greek, but rather from a Latinisation of the name of a mathematician living in Baghdad in the 9th century: Muhammad Ibn Musa Al-Khwarizmi. Latin translations of Al-Khwarizmi's algebra manual had circulated throughout Western Europe as early as the 12th century, the first being the Cambridge manuscript Dixit Algorizmi. The original Arabian manuscript has been lost. Through distortion, al-Khuwārizmī thus became algorizmi and algoritmi, then algorithm. On the history of these texts, see André Allard's reference edition, Muhammad Ibn Musa Al-Khwarizmi, Le calcul Indien (algorismus). Versions latines du XIIe siècle, Librairie scientifique et technique Albert Blanchard, Paris, 1992.

of social acceptability of an "automatic" decision, the final decision (is a human being at the end of the decision-making chain?), or responsibility in the event of a mistake. And, clearly, these issues are being raised again with the latest algorithmic innovations (see section 3 for a general presentation of artificial intelligence).

What is different, then, about the latest technologies that fall under the AI acronym? From an objective standpoint, changes include the quantity of information that can be handled by computers (big data) and the complexity of learning algorithms which, by feeding off big data, can perform perceptive and cognitive tasks that enable visual or audio recognition and make decisions in specific contexts. By combining different features (facial recognition, behaviour analysis, decision-making), AI raises extremely important ethical issues. From a subjective standpoint, what is new is the wake-up call to citizens, however late and suddenly this occurred, on issues of algorithmic governance, handling of personal data and the social impact already felt by some professional sectors.

If the progress of AI can surprise and fascinate, it can also evoke the fear that using machines, namely robots, will greatly diminish the aspect of human relationships when it comes to medical treatment, elderly care, legal representation or even teaching. Reactions to the development of artificial intelligence can even be hostile when AI is associated with increased control of individuals and society, a loss of independence and a curtailing of civil liberties. For this reason, a dark cloud always hangs over the hope that artificial intelligence will usher in social progress: placed in the wrong hands, AI could become a *weapon of mass domination* (control of private life, concentration of capital, new discrimination). Many people also question the intentions of researchers, developers, entrepreneurs and policymakers.

The development of AI and its applications therefore involve conflicting fundamental ethical values and create serious moral dilemmas and deep social and political controversies: should we promote public safety by increasing smart surveillance (facial recognition, anticipating violent behaviour) at the expense of individual freedoms? Can we objectively improve the well-being of individuals, namely by encouraging people to adopt behaviours normalized by smart devices (nutritional behaviour, work management, day planner) while still respecting people's independence? Should economic performance targets take priority over a concern for an equitable share of the benefits of the AI market?

These dilemmas or tensions cannot simply be resolved by ranking fundamental values and interests. Otherwise stated: it is not about classifying values in order of importance *a priori*, or building a simple and unequivocal scale of values, let alone promoting some while ignoring others (security at the expense of liberty, efficiency without social justice, well-being at the expense of independence). We also cannot aspire to find unique and permanent solutions. What we need to do is seriously contemplate the moral dilemmas caused by the development of AI and build an ethical, political and legal framework together that will allow us to deal with this while respecting the different fundamental values that we legitimately hold as members of a democratic society.

[2] Paul Meehl, Clinical versus Statistical Prediction, University of Minnesota, 1954.

2.2

FORUM ON THE SOCIALLY RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

These discussions were the starting point for an initiative by the Fonds de recherche du Québec and the Université de Montréal to organize an international meeting to discuss the social impacts of AI. Within this context, the Université de Montréal organizing committee suggested launching the work around the *Montréal Declaration for a Responsible Development of Artificial Intelligence* based on a consultative and participatory process[4]. On November 2 and 3, 2017, a forum on the ethical development of AI brought together leading experts in fields ranging from pure sciences to social sciences and humanities at the Palais des congrès de Montréal. The Forum proposed that guidelines for a collective reflection on the ethical and socially responsible development of artificial intelligence be established, with the following three objectives in mind:

- offer a public forum for dialogue on AI development issues and their social impact
- spark interest and raise visibility among decision-makers, industry partners, politicians and the general community interested in AI while bringing attention to social issues raised by the sudden growth and numerous uses of AI
- encourage an interdisciplinary and inter-industry approach as a key component to successful ethical and sustainable AI

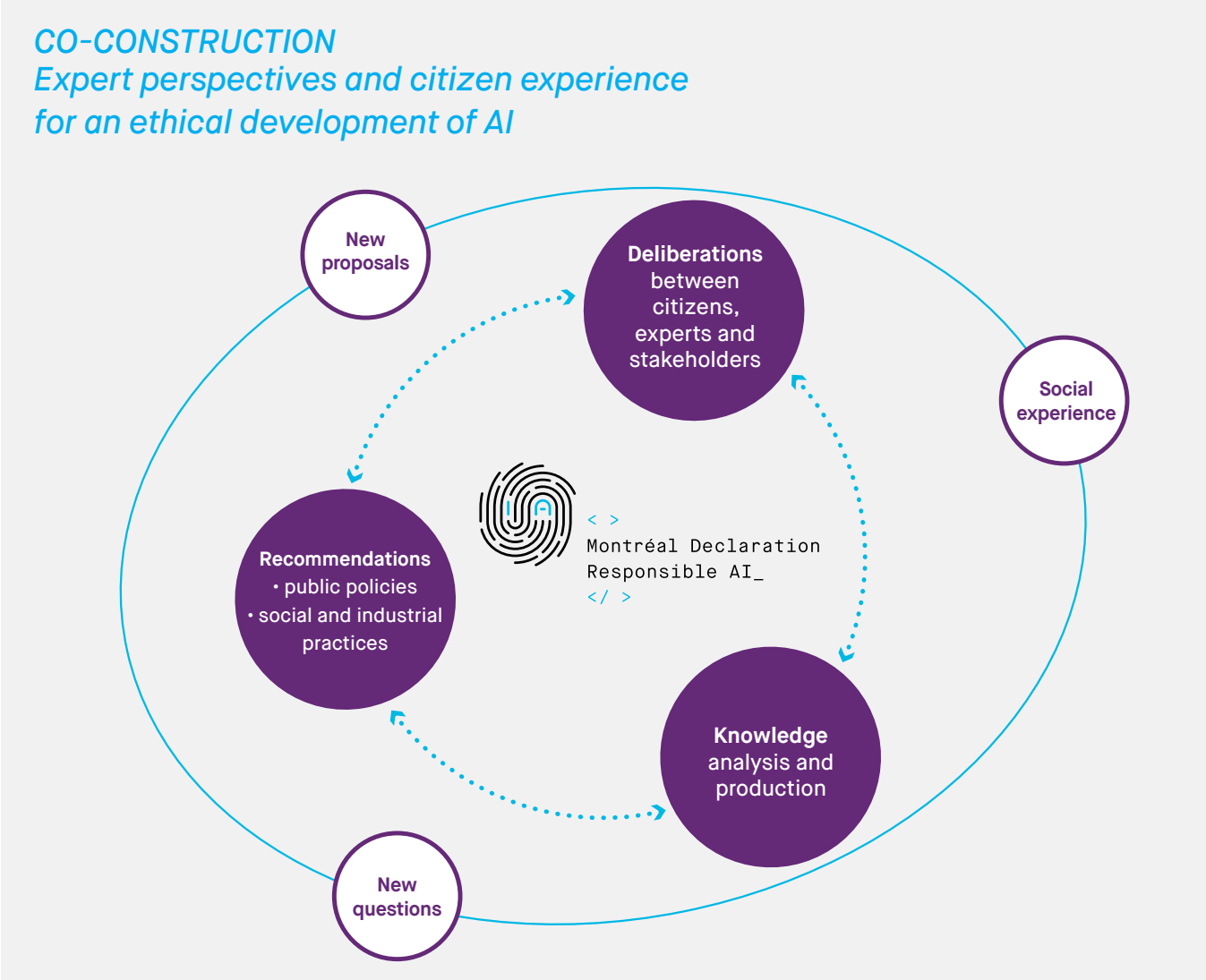
That is how the guidelines were defined on an inclusive (interdisciplinary and inter-industry) approach, which is key to developing the *Montréal Declaration* for a Responsible Development of Artificial Intelligence that is responsible, socially progressive and equitable and promotes social justice. The preliminary version of the *Montréal Declaration* was presented at the end of the forum. It was then a matter of launching the citizen co-construction process on AI ethics, a process we will expand upon in section 33.

[4] The Forum's scientific committee was made up of Louise Béliveau (Université de Montréal, Vice-rectorat aux affaires étudiantes et aux études), Yoshua Bengio (Université de Montréal, Département d'informatique, MILA, IVADO), David Décary-Hétu (Université de Montréal, École de criminologie), Nathalie De Marcellis-Warin (École Polytechnique, Département de mathématiques et de génie industriel, CIRANO – Centre interuniversitaire de recherche en analyse des organisation), Marc-Antoine Dilhac (Université de Montréal, Département de philosophie, CRÉ Centre de recherche en éthique), Marie-Josée Hébert (Université de Montréal, Vice-rectorat à la recherche, à la découverte, à la création et à l'innovation), Gregor Murray (Université de Montréal, École de relations industrielles et CRIMT – Centre de recherche interuniversitaire sur la mondialisation et le travail), Doina Precup (Université McGill, School of Computer Science; MILA), Catherine Régis (Université de Montréal, Faculté de droit, CRDP – Centre de recherche en droit public), Christine Tappolet (Université de Montréal, Département de philosophie et CRÉ – Centre de recherche en éthique).

2.3

TOWARDS THE MONTRÉAL DECLARATION FOR RESPONSIBLE AI DEVELOPMENT

Figure 3: The co-construction approach



As mentioned at the beginning of the chapter, the initial work of identifying these values and corresponding principles was conducted solely to launch the citizen participation process so that the ethical principles of responsible AI development could be refined, added to and completed. It should come as no surprise, then, that the preliminary version of the Declaration is schematic and that the statement of principles is intentionally very simple and consensus-based, so that they could be interpreted and completed during the public deliberations[5]. One year later, the *Declaration* has been considerably improved.

[5] The scientific committee in charge of writing this preliminary version was made up of Yoshua Bengio (Université de Montréal, Département d'informatique, MILA, IVADO), Guillaume Chicoisne (IVADO), Marc-Antoine Dilhac (Université de Montréal, Département de philosophie, CRÉ Centre de recherche en éthique), Vincent Gautrais (Université de Montréal, Faculté de droit, CRDP – Centre de recherche en droit public), Martin Gibert (CRÉ – Centre de recherche en éthique, IVADO), Pascale Lehoux (Université de Montréal, ESPUM – Ecole de santé publique), Joëlle Pineau (Université McGill, School of Computer Science; MILA), Peter Railton (Université du Michigan, , Académie américaine des arts et des sciences, philosophie), Christine Tappolet (Université de Montréal, Département de philosophie et CRÉ – Centre de recherche en éthique).

If one of the goals of the co-construction process was to refine the ethical principles proposed in the preliminary version of the *Montréal Declaration*, another equally important goal was to develop recommendations for overseeing AI research and its industrial and technological development. However, all too frequently we see analysis reports and recommendations forgotten as soon as they are published: that is why we must keep the momentum going during the co-construction period.

Once the co-construction process is complete (or suspended), we need to open public debate in forums where political, legal and policy decisions are made, so that recommendations from citizen deliberations may be concretely implemented. These recommendations are not simply legal in nature and, when they are, do not necessarily involve changing a law. They could, however, demand that the legal framework be modified; in some areas, they must. In other instances, the purpose of the recommendations is to feed and guide discussions held by professional organizations so that they modify their code of ethics or so that companies adopt a new ethical framework.

This step is the ultimate goal of the co-construction process. We must, however, immediately clarify that, when faced with a technology that has not stopped evolving over the past 70 years and whose major innovations arise every two to five years on average, it would be unreasonable to present the Declaration as definitive and complete. We need to think of co-construction as an open process, with successive and cyclical stages of deliberation, participation and recommendations, and see the Declaration itself as a road map that can be reviewed and adapted as AI knowledge and techniques evolve. This process of knowledge production, citizen deliberations and ethical framework and public policy recommendations will need to be expanded to a lasting institutional structure that allows it to respond to the evolution of AI.

2.4

MONTRÉAL AND THE INTERNATIONAL CONTEXT

The *Montréal Declaration* initiative is part of a dynamic scientific, social and industrial context. Montréal is a major hub for research and development in artificial intelligence, boasting a community of researchers, world-renowned university labs (MILA, IVADO) and an incubator full of thriving start-ups and businesses. This scientific, technological and industrial development is at the heart of a revolution that is transforming social practices, business models and lifestyles, and affecting all sectors of society. Thanks to its Laboratoire de l'innovation urbaine de Montréal, the Ville de Montréal is also a living lab of social and technological change[6]. With fundamental scientific research come social and ethical responsibilities that the Montreal AI community fully embraces.

But outside Montréal, Quebec and Canada also offer a social context that is conducive to reflecting on the social impact of AI. Like MILA in Montréal, Vector in Toronto, AMII (Alberta Machine Intelligence Institute) in Edmonton, and the CRDM (Centre de recherche en données massives) in Quebec are hubs of excellence in fundamental research that have brought about incredibly quick and robust industrial growth. The Canadian Institute for Advanced Research (CIFAR, or ICRA, a partner in the Declaration project) has played a leading role in the Canadian development of AI by supporting fundamental research when AI was going through its "winter". The Declaration initiative is supported by various stakeholders in Québec and Canada outside of Montréal.

Many international partners have also shown their support for the *Montréal Declaration*, especially its methodology. The Declaration team was able to establish a dialogue with institutions such as the Royal Society of the United Kingdom[7], the EGE (*European Group on Ethics in Science and New Technologies*[8]) and the European Commission's HLEG (*High Level Expert Group on AI*[9]), which have

[6] http://ville.montreal.qc.ca/portal/page?_pageid=5798,141982209&_dad=portal&_schema=PORTAL

[7] We wish to thank *UK Science and Innovation Network in Canada* who facilitated the dialogue.

[8] The European Group on Ethics in Science and New Technologies (EGE) is an independent advisory body of the President of the European Commissions.

their own study program and recommendations on AI. Immediately, we note similarities in the guidelines for ethical AI development as well as a shared desire to promote a democratic notion of AI use for the common good.

The *Montréal Declaration* initiative must also be viewed in the international context of an **AI spring.** The numerous initiatives that came before it must be highlighted because they acted as a catalyst for discussion on responsible AI. First, the Future of Life Institute, which was created in 2014, drafted the Asilomar Declaration in 2017: following a three-day conference, a declaration containing 23 fundamental principles on AI research and its uses was signed by more than 1,200 researchers. Professor Yoshua Bengio took part in the event and brought attention to the risks of irresponsible and malicious AI use[10].

Since the Asilomar Conference, there have been many reports published on AI ethics. The report from the Institute of Electrical and Electronics Engineers (IEEE), Ethically aligned design V2, was made public at the end of 2017 and brought together several hundred AI researchers and engineers. The AI Now Institute based in New York University has also produced several reports, the latest of which deals with evaluating the impact of AI[11]. Two ambitious strategic reports were published in March and April of 2018: the Mission Villani report in France and "AI in the UK: ready, willing, and able?" from the United Kingdom House of Lords. We must also highlight the participative approach of the CNIL (Commission nationale de l'informatique et des libertés) in France that led to the publication of a report with the evocative title, "How can humans keep the upper hand? - The ethical matters raised by algorithms and artificial intelligence", in December 2017.

How does the *Montréal Declaration* position itself among these many independent initiatives? And what about the rise of ethical issues in AI? This last question so important that we include the same warning as the EGE in its report Artificial Intelligence, Robotics and "Autonomous" Systems (March 2018) that in the absence of coordinated reflection on the ethical and social issues of AI, a risk of "ethics shopping"[12] exists. The immediate consequence is a sort of "offshoring" of ethical costs in areas of the world where ethical criteria are low priorities. Another risk is trivializing ethical discourse.

Each step in developing the ethical framework has merit. Part 2 of this report provides a "Overview of international AI ethics recommendations for 2018". The *Montréal Declaration* initiative is different in that it is essentially participative. From February to November 2018, the co-construction process brought together over 500 citizens, experts and stakeholders during 15 workshops, co-construction days and roundtables across Québec and Europe. Although other participative initiatives have been led elsewhere, namely in France, the *Montréal Declaration* is unique in both its size and foresight methods.

The *Montréal Declaration*'s vocation is to open a forum for dialogue in Québec and Canada and offers a platform for a collective think tank that extends beyond Canadian borders. The goal is to identify socially acceptable and innovative AI trends using informed citizen discussions as a benchmark for the different democratic societies concerned. Citizens of non-democratic societies who wish to take part in a global debate on the future of human societies must also have access to this forum for dialogue.

[9] The HLEG on AI is a group of 52 experts selected by the European Commission to define the application principles of Europe's AI strategy. We thank the people in charge of the HLEG for allowing us to take part in their work between September and November 2018, in order to share and enrich our respective reflections and experiences.

[10] Yoshua Bengio interview during the Asilomar conference: futureoflife.org/2017/01/18/yoshua-bengio-interview/

[11] AI Now Institute, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability", April 2018.

[12] EGE, *Artificial Intelligence, Robotics and 'Autonomous' Systems* (March 2018), p. 14.

3. THE ETHICAL AND SOCIAL ISSUES OF AI

The collective reflection process at the heart of the development of the *Montréal Declaration* is based on a preliminary version of the Declaration of ethical principles itself and informative exposés on AI and the ethics of AI.

3.1

WHAT IS AI?

The idea of AI is not new. As early as the 17th century, philosopher and mathematician Leibniz came up with the idea of a universal characteristic and combinatorial art: reasoning comes down to calculating, and thought is conceived in algorithmic fashion[13]. The notion of calculus ratiocination (logical calculation) predates the idea of an intelligent machine as it was developed three centuries later in the 1940s by Alan Turing. In a 1948 report entitled "*Intelligent Machinery*" and in 1950, in his famous article "*Computing Machinery and Intelligence*"[14], Turing discusses a machine's intelligence and develops the imitation game to define the conditions in which a machine can be said to think. The term artificial intelligence appears for the first time in 1955 in the description of a workshop offered by John McCarthy (Dartmouth College), "2-month, 10-man study of artificial intelligence". But the uses and development possibilities seemed very limited then, and so began the AI winter, with minimal interest from the scientific community. Yet, if the discipline's development paled in comparison to the philosophical and cultural fervour it inspired (one need only recall *2001: A Space Odyssey, Blade Runner* or *Terminator*, to name but a few hit movies), research in the field never ceased, and the dawn of the 21st century ushered in an AI spring.

In a certain way , AI consists of simulating human intelligence[15], drawing inspiration from it and reproducing it. But, above all, it is the brain, the human intelligence headquarters, which was designed as a machine capable of gathering, identifying and collecting data from its environment that it can then analyze, interpret and understand, using this experience to establish connections. The field of AI research consists of producing mathematical tools to formalize how the mind operates, thereby creating machines that can perform more or less basic cognitive tasks associated with natural human intelligence. For example, recognizing complex patterns among a large quantity of data, or reasoning in probabilistic fashion to classify information according to categories, predict quantitative data or group data together. These cognitive skills are the basis for other skills such as deciding among many possible actions to achieve a goal, interpreting an image or sound, predicting behaviour, anticipating an event, diagnosing a condition and so forth.

But these cognitive skills can only exist if the machine is also capable of identifying sensitive shapes such as images and sounds, which has been made possible by recent computer innovations. The notion of AI, therefore, also encompasses visual or sound recognition technologies that allow the machine to perceive its environment and construct a rendering of this environment.

Two elements underpin the achievements of AI: data and algorithms, meaning a series of instructions that perform a complex action. Simply put, if you want to cook a new dish, you need to know the ingredients (the data) and follow a recipe that provides instructions on how to use them correctly (the algorithm). Up until now, data processing capacities

[13] Leibniz (1666), *De Arte combinatoria*.

[14] A. M. Turing (1950), "Computing Machinery and Intelligence". Mind 49, p. 433-460.

[15] Alan Turing begins his "Intelligent Machinery" (1948) report as follows: "I propose to investigate the question as to whether it is possible for machinery to show intelligent behaviour."

(quantity of data and processing algorithms) were too limited to imagine a useful development for AI technologies. Things changed with the use of materials that made it possible to build very small and very fast calculators (computer chips) and store massive amounts of data as well as the dawn of the information era with the Internet.

What changed is the gigantic amount of data we can not only generate and transmit, but also process. If big data existed in the past, for example in the financial industry, nowadays it is a multitude of inanimate objects, spaces or receivers that are constantly producing unstructured or structured data, which must be manipulated and transformed before it can be used (data mining). It can be millions of messages published on social media, all the words contained in a library full of thousands of books, or content from huge image banks.

But what changed also is the type of algorithm developed by AI researchers. Determinist algorithms, which are a determined set of instructions like a cooking recipe, are being replaced by learning algorithms which rely on increasingly complex neural networks as the calculating power of machines increases. In computing, we talk about machine learning, and the progress of this field of study was reinforced by the development of deep learning. At the heart of the notion of AI itself is the ability to adapt and learn. In fact, for a machine to be considered intelligent, it must be able to learn by itself from the data it receives, as a human being does. And just like with humans, machine learning can be supervised, or not, by human beings that train machines on data.

It is these deep learning techniques that allowed machines to surpass human beings in complex games such as chess with AlphaZero, which also beats any other machine that doesn't use deep learning, and the game of Go, which was reputedly unbeatable at algorithms, but which saw AlphaGo triumph over the best players in the world in 2015.

Although these examples are telling, AI can also serve other purposes such as automating tasks that require human intervention, especially tasks such as perception and recognition. For example, processing speech, recognizing objects, words, shapes and text, interpreting scenes, colours, similarities or differences in large sets, and by extension analyzing data and decision-making—or help with decision-making. The possibilities are incredibly vast, and increase tenfold every time engineers and programmers combine them to create new uses.

3.2

AI IN EVERYDAY LIFE AND PHILOSOPHICAL QUESTIONING

AI engages us in an ethical reflection that, unlike one concerning nuclear or genomics, deals with everyday objects and technologies. AI is all around us and shapes our lives more than ever. We are used to wearing small connected objects (phones, watches) and we are preparing for the arrival of self-driving vehicles, cars and buses, but already we take trains and subways that operate independently, and planes on autopilot can take off, steer and land without human intervention. We use ranking algorithms for our Internet searches, autocorrect built into our messaging apps, curation apps for music or meetups, and we know that companies use sorting algorithms, banks use management and financial investment algorithms, and that certain medical diagnoses can now be very exactly made by algorithms, and the list goes on.

These technologies are so seamlessly integrated into our everyday life that we no longer really think about them. When we talk about AI, most people still associate it with menacing, multifunctional machines that have some sort of consciousness, able to formulate a plan to destroy all humans[16]. Yet the AI experience is a thoroughly banal one nowadays, with recommendation algorithms flooding the Internet (Google, Amazon, Facebook). If you're shopping online, there's a good chance a pop-up window will open and that Inès will start up a conversation with:

"Hi, how can I help you shop today?"

"Hi Inès"

[16] Stanley Kubrick masterfully captured (and helped craft) this fantasy with the very human computer HAL 9000, in his film *2001: A Space Odyssey* (1968).

For a few moments, you have the impression that there is someone named Inès behind the screen talking to you; for a few moments, it's okay to wonder. Inès asks you questions, answers yours, provides the important information you need to continue shopping. But after a little back-and-forth, you realize that although Inès provides relevant information, she replies in mechanical fashion, she doesn't understand the way you write, doesn't get jokes or open-ended questions, in other words, she doesn't interact naturally with you. Inès is a conversational agent, a chatbot, AI. It has become commonplace to chat online with chatbots to get more information about your health plan or new bank account, or even fashion advice.

For now, chatbots can be spotted within a few minutes of conversation, usually much sooner. If a chatbot could go undetected by a human being for a reasonable amount of time, we would consider that the machine successfully passed the Turing test and we would, according to this test, be dealing with a case of artificial intelligence, meaning a machine that thinks.

In his famous article, "Computing Machinery and Intelligence", the father of modern computing, Alan Turing, proposes an answer to the question: "Can a machine think?"[17] And yet, in the introduction of his article, he changes the problem he feels he can provide an answer to: can a machine act in such a way that it is indistinguishable from a human being? He then offers the famous "imitation game", which consists of putting a human being that asks questions (the interrogator) in contact with another human being and a machine answering his questions. If the machine can imitate a human being to the point where the interrogator cannot tell whether the human being or the machine replied, we can conclude that the machine thinks. This is what is meant by the "Turing test".

This imitation game caused a great deal of controversy and saw philosophers fiercely opposed over whether a machine could be said to think. An experiment known as the "the Chinese chamber" was made popular in the 1980s by philosopher John Searle[18]. According to Searle, a machine that outwardly acts in the same fashion as a human being cannot be considered to have intelligence in the true sense of the word. To illustrate this point, Searle asks us to imagine a room in which a person who, knowing nothing of Chinese, will try to pass for a Chinese speaker. It's a variation of the imitation game: the person in the Chinese room, let's call him John, receives messages written in Chinese that Chinese speakers outside the room hand him. John doesn't understand a word of the messages he receives, but he possesses a very complex instruction manual which allows him to manipulate the Chinese characters and compose replies that are understood by Chinese speakers outside the room, so that they believe that the reply was written by someone who speaks Chinese. Searle deducts that in this case John simulated language skills but doesn't possess them; he made people believe he understood Chinese, but he didn't understand what he was writing. According to Searle, the same conclusion goes for AI: an intelligent machine manipulates characters, it follows an algorithm, meaning a series of instructions to accomplish a task (in this case, write), but doesn't understand what it's doing.

The debate is a fascinating one and far from being settled, but we don't really need to answer Turing's question to consider the place AI holds in our lives and in our societies. For now, well-trained chatbots can converse as well as humans within a very limited framework of conversations, but leave no one guessing once that framework changes. And even if AI is ushering in an era where it is harder and harder to tell a naturally intelligent being from an artificially intelligent one, intelligent machines remain tools developed to accomplish well-defined tasks. We can, then, leave it up to cognitive philosophy metaphysics, psychology and neuroscience to debate the concept of artificial intelligence and discuss the possibility of robots developing emotions and feeling empathy[19]. The questions that arise with the introduction of AI in our lives are of a practical nature, whether ethical, political or legal. It is a questioning of the values and ethical principles, public policy orientations and applying standards surrounding AI research and its uses.

[17] A. M. Turing (1950).

[18] J. Searle (1980), 'Minds, Brains and Programs'. *Behavioral and Brain Sciences* 3, p. 417–57.

[19] Which is very different from questions on the use of machines to detect human emotions, process them and answer in adequate fashion. See for example the work of Rosalind W. Picard, *Affective Computing*, Cambridge, MIT Press, 1997.

Because AI technologies are indifferent to their multiple uses, the problem is not knowing whether AI is good or bad in and of itself, but of determining which uses and goals are ethical, socially responsible, and compatible with democratic values and political principles. However, this ethical reflection does not only concern the uses of AI, but also AI research, its general orientations and goals. Nuclear research was not initially destined to produce bombs with tragically powerful consequences for humanity. Many scientific programs, however, did have that goal. We must, then, pay close attention to the direction AI research takes, in universities and as well as private corporations or government organizations.

3.3

THE ETHICAL ISSUES OF AI

Why introduce ethics when we can discuss the societal, social and economic impacts of AI? Can we afford the luxury of an ethical reflection? And isn't it a bit naive to want to provide an ethical framework for AI development, which generates colossal profits? These are questions ethicists hear on a regular basis among sceptical citizens, as well as decision-makers who understand the extent of their leeway. To answer this, we must first briefly present the field of ethics when discussing the social issues of AI.

Simply put, ethics is a reflection on the values and principles that underlie our actions and decisions, when they affect the legitimate interests of other people. This supposes that everyone can agree on a person's legitimate interests, and this is precisely what feeds the debate in ethics. The field of ethics is therefore not concerned with what can be done, but generally what must, or should be done: we can kill a million people with a single nuclear bomb, but must it be done to impress an enemy country and demoralize a population already suffering from war? Take a less tragic example: you can lie to a friend about their new haircut, but is it moral in order to save them from being hurt? What must be done in this case? To answer that question, we must examine the available options: tell the truth, or not tell it, or tell only part of it, or tell it in a certain way. We must also examine the consequences of each option, question if they are important, and if so, why. We must also reflect on the objectives which are valorous (doing good unto others, respecting others). Finally, we must give ourselves a rule, a moral principle: for example, the categorical principle according to which it is always wrong to lie, regardless of the consequences; or the hypothetical principle according to which it is not morally right to lie unless...

The field of ethics that applies to AI issues is public ethics. If we use the same type of reflection as public ethics, the subject isn't the same, nor is the context for reflection. Public ethics is concerned with all questions that involve difficult collective choices on controversial institutional and social practices that affect all individuals as members of society, and not as members of a particular group: should a doctor tell his patient the truth about his health condition even if it will depress him and speed up the progress of the disease? This question doesn't concern the doctor's personal morality, but the types of behaviour we can rightfully expect from someone who holds the social role of doctor. This question is of a public nature and should be the subject of a public debate to define, using social values, best practices in terms of the patient-doctor relationship. By public debate, we mean all types of discussions which can take many forms of consultations, deliberations or democratic participation, and which are open to diverse individual and institutional stakeholders such as professionals in the field, association or union representatives, experts, policymakers and citizens. Public ethics calls for a collective reflection to establish best practice principles and demands that the stakeholders justify their suggestions based on acceptable arguments in a context of pluralism. In the case of the medical lie, you can appeal to shared values such as independence, respect for the person, dignity, the patient's health or well-being, etc. Out of these values, principles can be established that guide the practice of medicine and provide paths to regulation by implementing a code

[18] J. Searle (1980), 'Minds, Brains and Programs'. *Behavioral and Brain Sciences 3*, p. 417–57.

[19] Ce qui est très différent des questions sur l'usage des machines pour détecter les émotions humaines, les traiter et y répondre de manière adéquate. Voir par exemple les travaux de Rosalind W. Picard, *Affective Computing*, Cambridge, MIT Press, 1997.

of ethics, modifying a law or introducing a new law.

Public ethics are not beside nor above the law, which has its own logic, but help clarify the issues of social life that various stakeholders must keep in mind to meet citizens' standard expectations and ensure equitable social cooperation. In this sense, public ethics shape public policies, and can lead to legislation, regulation, a code of ethics, an audit mechanism and more.

In AI, it is this kind of ethical reflection that we introduce. Let's take the example of Melody, a medical conversation agent. Melody makes online diagnoses that you can access on your cellphone according to the symptoms you describe. In a certain way, it acts like a doctor. This can be very practical in a society where the healthcare system is either inaccessible or underdeveloped. But the fact that it is practical is not sufficient to authorize the public release of an app like Melody. Indeed, this app raises ethical questions that were not readily apparent with Inès, the shopping advisor chatbot. For example, we need to debate whether Melody must give users every possible prognosis, even if they are not equipped to understand the information. This problem simply transposes ethical medical questioning which has already received a normative response for which there is widespread consensus. The notion of informed consent, of a patient's free and enlightened decision helps clarify a doctor's obligations. Does this solve the problem that Melody and its sister applications often multiply unchecked?[20] Overall, probably, but when specific attention is paid to this technology it is not that simple. The context does not allow Melody to ensure that the patient understands the diagnosis, or the urgency or not of treating the diagnosed condition. What rules must be invented to guarantee a patient's autonomy and well-being? That is the issue of collective deliberation on AI's ethical issues.

Ethical solutions have yet to be found for other issues specific to AI. For example, if Melody makes a wrong diagnosis, and the condition of the user who followed her advice takes a serious turn for the worse, who is responsible? In the case of a medical consultation with a human doctor, it is very easy to determine who is responsible for a medical error, but that is not the case with decision-making algorithms. Do you hold the algorithm responsible? The developer, or rather the company that developed the algorithm and that makes money from its use? And if the product is certified, shouldn't the certifying body be blamed and penalized?

Public ethics questioning forces us to think about institutions that offer credible responses to a moral dilemma. It also deals with the type of society we want and the principles of its organization. By pursuing the reflection on medical chatbots, the use for developing such intelligent machines, from a social and human standpoint, is undeniable. We must question, then, whether it is acceptable for smart apps to replace medical doctors, assuming they can make an accurate diagnosis, even more accurate than a human. What does a patient-doctor relationship look like when the doctor is a chatbot? What essential elements are gained and which are lost? This is not a "utilitarian" type of question, but a question about the importance of our social relationships, recognizing our vulnerability as patients, our human identity. Let's take it one step further: investing in the development of this type of AI rests on an eminently arguable social choice, which requires a collective discussion on the type of society we wish to build. We can consider the need for improving access to an efficient public healthcare system and therefore further invest in medical training and an equitable health organization.

3.4

AI ETHICS AND THE MONTRÉAL DECLARATION

The development of AI and its uses, then, involves fundamental and conflicting moral values that can provoke serious ethical, social and political controversies: should we develop apps like Melody to diagnose isolated people more quickly, or invest differently in the healthcare system so everyone can see a human doctor? There is no simple answer, but choices must be made.

[20] The British public health service, the NHS (*National Health Service*) recently created a library of trustworthy apps (*NHS Apps Library*). Apps that do not offer sufficient guarantees can be deleted from the library, which brings serious commercial repercussions for the company selling the app.

The *Montréal Declaration* provides a basic moral vocabulary to identify, analyze and form practical answers to problematic social situations. The analysis of the Melody chatbot case illustrates the purpose of the Declaration. To understand the issue of enlightened patient understanding of a diagnosis, attributing fault in the event of a wrong diagnosis or accessing health services, the *Montréal Declaration* offers a list of values you can immediately consult: autonomy, responsibility, equity or justice. The principle of privacy, for example, helps frame the problem of patient data confidentiality.

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

1. Although they are presented as a list, there is no hierarchy. The last principle is no less important than the first. However, depending on the circumstances, it is possible to lend more weight to one principle than another, or to consider one principle more relevant than another.

2. Although they are diverse, they must be subject to a coherent interpretation to avoid 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.

3. Although they reflect the moral and political culture of the society in which they were developed, they comprise the basis for an intercultural and international dialogue.

4. Although they can be interpreted in different ways, they cannot be interpreted in just any way. It is imperative that their interpretation be coherent.

5. Although these are ethical principles, they can be translated into political language and interpreted in legal fashion.

The Declaration of principles is followed by a list of recommendations that act as guidelines for the digital transition within the Declaration's ethical framework. This list does not aspire to be exhaustive, nor can it cover every aspect of AI application; such ambition would be doomed. Rather, it aims to address a few key cross-sector themes so that we can think about how to make the transition towards a society in which AI helps promote the common good: algorithmic governance, digital literacy, digital inclusion of diversity and environmental sustainability.

The *Montréal Declaration* was designed for any person, organization and company that wishes to take part in the responsible development of artificial intelligence, whether to contribute scientifically or technologically, develop social projects, establish rules (regulations, codes) that apply to it, contest harmful or unwise approaches, or alert public opinion when necessary.

It is also designed for political representatives, whether elected or appointed, whose citizens expect them to respond to developing social changes, quickly establish a framework that encourages digital transition for the greater good, and anticipate serious risks presented by AI development.

The recommendations that follow the Declaration are intended more specifically for stakeholders in AI development in Quebec and Canada. They are examples of concrete measures collectively developed out of the Declaration's ethical considerations. For this reason, they can act as a benchmark for stakeholders in AI development outside Canada.

4. THE CO-CONSTRUCTION APPROACH

4.1

THE PRINCIPLES OF THE CO-CONSTRUCTION APPROACH

To answer the many questions raised by the use of intelligent machines and ensure that AI develops "in an intelligent manner" within democratic societies, we need to solicit an "excess" of democracy and involve the greatest number of citizens in the reflection process on the social issues of AI. The goal of the co-construction approach is to open a democratic discussion on the way society must be organized to use AI responsibly.

It is not only a matter of knowing what people think of a certain innovation and surveying their "intuitive" preferences; co-construction is not a public opinion poll on questions such as: "Are you scared that AI will replace judges?", "Would you prefer that a human or a robot operate on you?" This is an interesting question, and the survey method provides important information to policymakers as well as important working material for social sciences. However, although co-construction invites people to think collectively about democratic issues, it also calls for well-argued answers to pressing questions to be developed and political and legal recommendations to be formulated. The co-construction process also lends them a certain democratic legitimacy, which creates the conditions for a political debate and accountability from policymakers, professionals and industry stakeholders.

This is the entire reasoning behind the approach initiated by the Montréal Declaration: entrust democratic societies with the responsibility of resolving moral and political issues that affect society as a whole. The future of AI is not only written in algorithms; it resides foremost in collective human intelligence.

4.1.1 The Principles of Good Citizen Involvement

The moment you involve the public in a consultation and participation process (co-construction) on controversial social questions, you must ensure that the process avoids the risks usually associated with a democratic exercise. And yet, two objections are traditionally brought up to discredit public involvement[21]:

1. Ignorance: according to this objection, the most common, the public is ignorant and cannot understand complex issues that require scientific knowledge, master logical forms of argument and understand political and legal processes.
2. Populism: according to this objection, involving an unqualified public creates an opportunity for demagogic manipulation that fuels popular stereotypes and can lead to unreasonable proposals being adopted that are hostile to social progress or tyrannical towards minorities.

We do not share the belief that the public is so ignorant that they must not be consulted. We do not subscribe to the idea that non-expert members of our society have unsurmountable prejudices and their alleged irrationality causes them to make systematic errors. Ignorance is certainly an important problem, but we believe instead that they can shed light on neglected aspects of social controversies, because they are concerned by the issues discussed, and they can contribute to significant solutions that experts haven not thought of, or were unable to support publicly.

If, for some individuals, prejudices and irrational tendencies cannot be completely eliminated, these biases can be overcome collectively. In favourable conditions, non-expert individuals can take part in complex debates on social issues, such as those presented today by artificial intelligence research and its industry applications. Experts in various matters relevant to citizen involvement on artificial intelligence can help implement these favourable conditions.

[21] Literature questioning the political ability of citizens have experienced a resurgence in recent years. See namely Jason Brennan, *Against Democracy*, Princeton, PUP, 2016; Ilya Somin, Democracy and Political Ignorance, Stanford, SUP, 2013.

We have identified four (4) conditions required for the co-construction process: epistemic diversity, access to relevant information, moderation and iteration.

A. EPISTEMIC DIVERSITY

We must first ensure that the deliberating groups are as diverse as possible, in terms of social environment, gender, generation and ethnic origin. This diversity is not only inherent to the idea we have of an inclusive democracy, but is required to increase the epistemic quality of the debates. This simply means that every person brings a different perspective to the subject being debated, and thus enriches the discussion[22].

B. ACCESS TO RELEVANT INFORMATION

We know, however, that epistemic diversity is insufficient and that if the participants have no skills or knowledge in the field being discussed, they cannot produce new knowledge, or follow the discussion. Collectively, then, they may increase individual errors. We must, therefore, prepare participants by providing relevant and quality information that is both accessible and reliable. An information session needs to be held prior to the deliberations.

C. MODERATION

Aside from having quality information, participants need to reason freely; that is, without being impeded by cognitive biases. We define cognitive bias as a distortion of rational thought by intuitive mechanisms. One of the most common and problematic biases in a deliberation is the confirmation bias: we tend to only accept opinions that confirm our own beliefs, and reject those that go against what we already believe. There are dozens of cognitive biases that can skew our logical train of thought.

But there are also biases that apply to the deliberation itself, such as the tendency to adopt more and more radical positions: if the group that is deliberating is initially distrustful of innovations in artificial intelligence, they will likely be entirely hostile towards them at the end of the deliberation. To avoid this type of knee-jerk reaction, we feel it is important to ensure epistemic diversity in the deliberating group and introduce a moderating body.

This does not necessarily have to be personal input from a moderator. Although we do not object to individual moderation, we believe we can overcome deliberation biases through other means, such as introducing unexpected events in scenarios that sparked the discussions.

D. ITERATION

Ideally, we should be able to bring together the entire population to reflect on the responsible development of artificial intelligence. But the conditions described above cannot be applied to very large groups, let alone a society of several million people. It is important, then, to involve citizens in smaller groups and increase the number of meetings. This is the iteration phase of co-construction.

[22] Estlund, David M. (2008). *Democratic Authority: A Philosophical Framework*. Princeton University Press.

The reasons to proceed this way are technical, but easily understood. The Marquis de Condorcet, a mathematician and key figure in the French Revolution, demonstrated that the judgment of groups is always more sound than each person individually, and that this increases as the group grows larger. For this to occur, however, two conditions must be met: the individuals in the group must have more than a fifty/fifty (50/50) chance of being right, and they must not communicate with one another (Condorcet rightly feared the risks of manipulation).

However, we cannot guarantee that for very large groups individuals will possess the required skills nor that each individual has more than a fifty-fifty chance of having an appropriate opinion. Encouraging deliberation (communication between one another) is one way of increasing the skill of participants, as long as it is done within the framework we are suggesting. Of course, that does not satisfy Condorcet's second condition, but it does guarantee the first. And to increase the quality of opinions, the number of deliberating groups must be multiplied: since we cannot increase the size of the group, we must increase the number of participants by conducting a series of participation sessions[23].

For these reasons, we prefer the structure of a co-construction workshop that brings together non-expert citizens, experts, stakeholders (associations, unions, professional representatives, businesses) as well as political representatives. These workshops are organized in different formats adapted to the deliberation spaces and satisfy the conditions for productive and robust citizen engagement. It is worth noting that the Declaration's roll-out is complex, and relies on other types of consultations: online surveys, reports and expert roundtables. The Declaration is a not simply a straightforward record of what was said in the co-construction workshops, but the result of multiple deliberations and reflections based on the co-construction workshops.

4.1.2 Experts and citizens

Why allow citizens to be heard on complex ethical and political questions that require a solid grasp of the technologies being discussed? Why not only consult the experts? There are many reasons, but the easiest is that AI affects everyone's lives; therefore, it concerns everyone and everyone must have a say in the socially desirable goals of its development.

Even when we are not, strictly speaking, faced with a dilemma, public ethics issues cannot be resolved without making choices that favour certain moral interests over others, while still not neglecting them. This is the result of pluralistic values that define the moral and political context of modern democratic societies. It is possible to promote well-being by challenging the priority of consent: think of a medical app that could access personal data without our consent, but that would help treat serious diseases thanks to the data.

This type of ethical and social choice should be in the hands of all members of our democratic society, and not just a part, a minority, even if they are experts.

The role of experts role is not to solve the ethical dilemmas raised by artificial intelligence themselves, nor become legislators. What are the experts doing then? The experts involved in the co-construction process of the Montréal Declaration have no intention of thinking for citizens, nor suggesting a legal and ethical framework that citizens would merely rubberstamp. Expertise must be used to support citizen reflection on complex social and ethical AI issues.

[23] Estlund (2008); Landemore, Hélène. (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press.

Sometimes ethicists can appear preachy, possess the answers to difficult questions that the public itself is still asking, or seem to be able to solve tomorrow's problems before they even happen. It is important to specify their role. Ethicists play three modest but crucial roles. They must:

> Ensure conditions exist that encourage citizen involvement

> Clarify ethical issues underlying the controversies around artificial intelligence

> Rationalize arguments being defended by participants by pointing out the arguments we know to be wrong or biased and explaining the reasons why they are wrong.

The role of ethicists is, therefore, to provide informed guidance[24]. Experts in other research fields (computer sciences, health, safety, etc.) also play a role by providing participants with the most useful and reliable information on the subject of controversy (How does an algorithm that learns to make a diagnosis work? Can a doctor be replaced by a robot programmed to make a diagnosis? What protective measures can we introduce to thwart attempts to hack our medical data?, and so forth.)

However, we should acknowledge that experts themselves often show important cognitive biases. They can be too optimistic or pessimistic about new technologies they know well; they also tend to lend too much weight to their own opinion, especially when they believe they can predict the evolution of their field of research and social change. By involving them as citizens in the co-construction workshops, we reduce the biases associated with expertise, as well as the role of authority created by the discrepancy in knowledge between them and other participants.

The co-construction workshops are forums for participation that help guide the socially desirable development of AI and innovate through proposals that shake up traditional analysis frameworks. The vital contributions from citizen deliberations are then analyzed and expanded upon by work committees composed of experts from different fields (researchers, professionals). The work of expanding and drafting recommendations follows the guidelines defined by the deliberation and stays true to the proposals issued at the co-construction workshops.

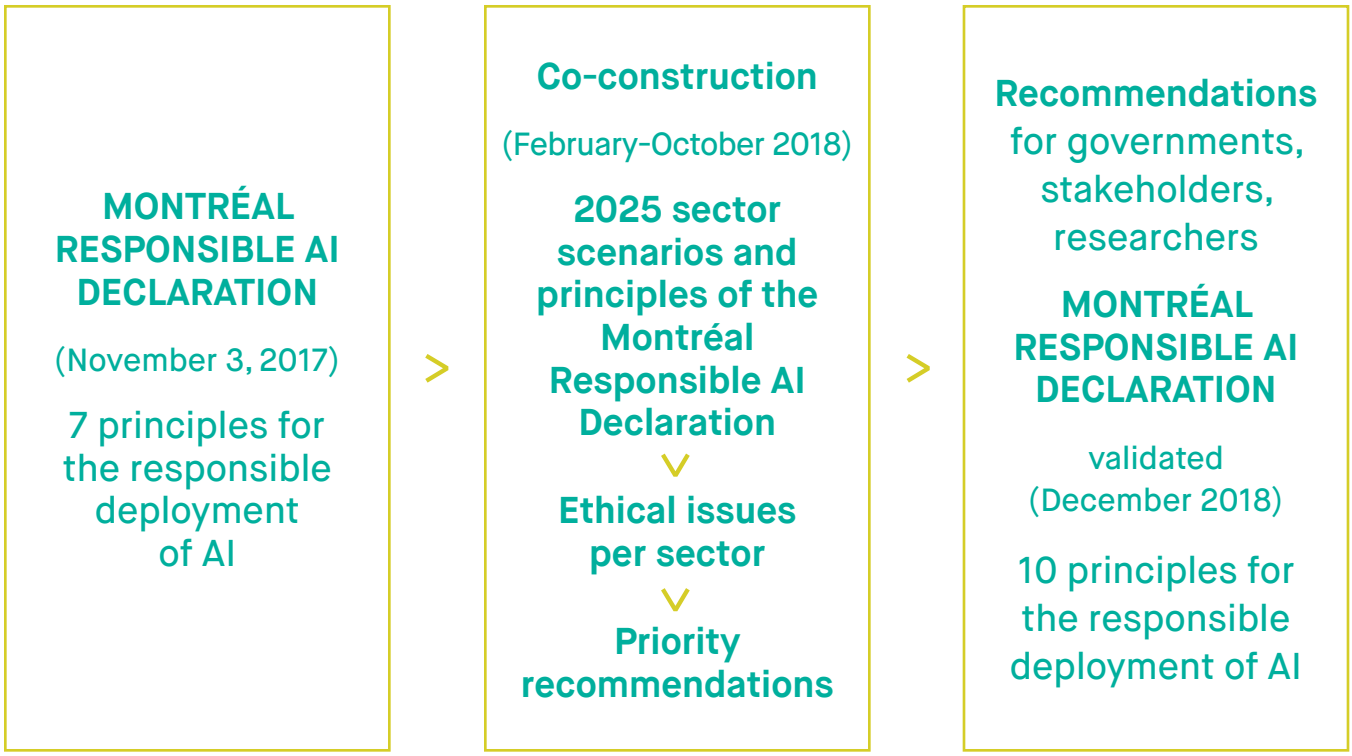
4.2

THE CO-CONSTRUCTION WORKSHOP METHODOLOGY

The first version of the Montréal declaration on Responsible AI, presented November 3, 2017, during the Responsable AI Forum, is the foundation for the co-construction process. Schematically, after having defined the "what"? ("which desirable ethical principles should be gathered in a declaration on the ethics of artificial intelligence"), the new phase becomes a matter of predicting—with citizens and stakeholders—how ethical controversies surrounding AI could surface in the next few years (in the fields of health, law, smart cities, education and culture, the workplace, public services), then imagining how they could be solved (for example, with a device such as sector certification, a new stakeholder/mediator, a form or standard, a public policy or research program).

The goal of the co-construction process and its workshops is primarily to test the principles of the Montréal Declaration for Responsible AI using foresight scenarios. Ultimately, the process will help specify ethical issues per sector, and establish priority recommendations for the AI community.

[24] Weinstock, Daniel M., *Profession éthicien*, Montréal, Presses de l'Université de Montréal, 2006.



More than ten co-construction workshops were held between February and October: three-hour world cafés in public libraries, and two important day-long co-construction workshops with various citizens, experts and stakeholders (at the SAT in Montréal, at the Musée de la civilisation in Québec City and the Centre Culturel Canadien in Paris[25]).

Choosing to organize world cafés in public libraries is directly linked to how these spaces are being reinvented to provide public services in Quebec and Canada[26]. By moving from a space that lends books to that of an inclusive "third space" that seeks to empower all citizens (e.g. with digital literacy services, citizen support, cultural mediation and discussion areas, lending tools and fab labs), public libraries will certainly play a key role in the responsible deployment of AI in Quebec and Canada.

The co-construction days were held in iconic spaces (Société des arts technologiques in Montréal, Musée de la civilisation in Québec) and focused primarily on uniting stakeholders and the very diverse disciplines that must work together to determine how AI should be responsibly deployed in society.

4.3

UNIQUENESS OF THE CO-CONSTRUCTION APPROACH

When compared with other AI ethics initiatives currently underway in the world, this co-construction process features three particularly original and innovative dimensions:

> Using foresight methods, with sector scenarios set in 2025 and the use of short stories to illustrate how ethical controversies on AI could surface or grow in the next few years (in the fields of health, law, smart cities, education and culture, the workplace). These 2025 scenarios, which present a variety of possible situations in a wide-open future, will be used to spark debate and

[25] We thank the Canadian Embassy in Paris for making the Paris workshop possible on October 9, 2018.

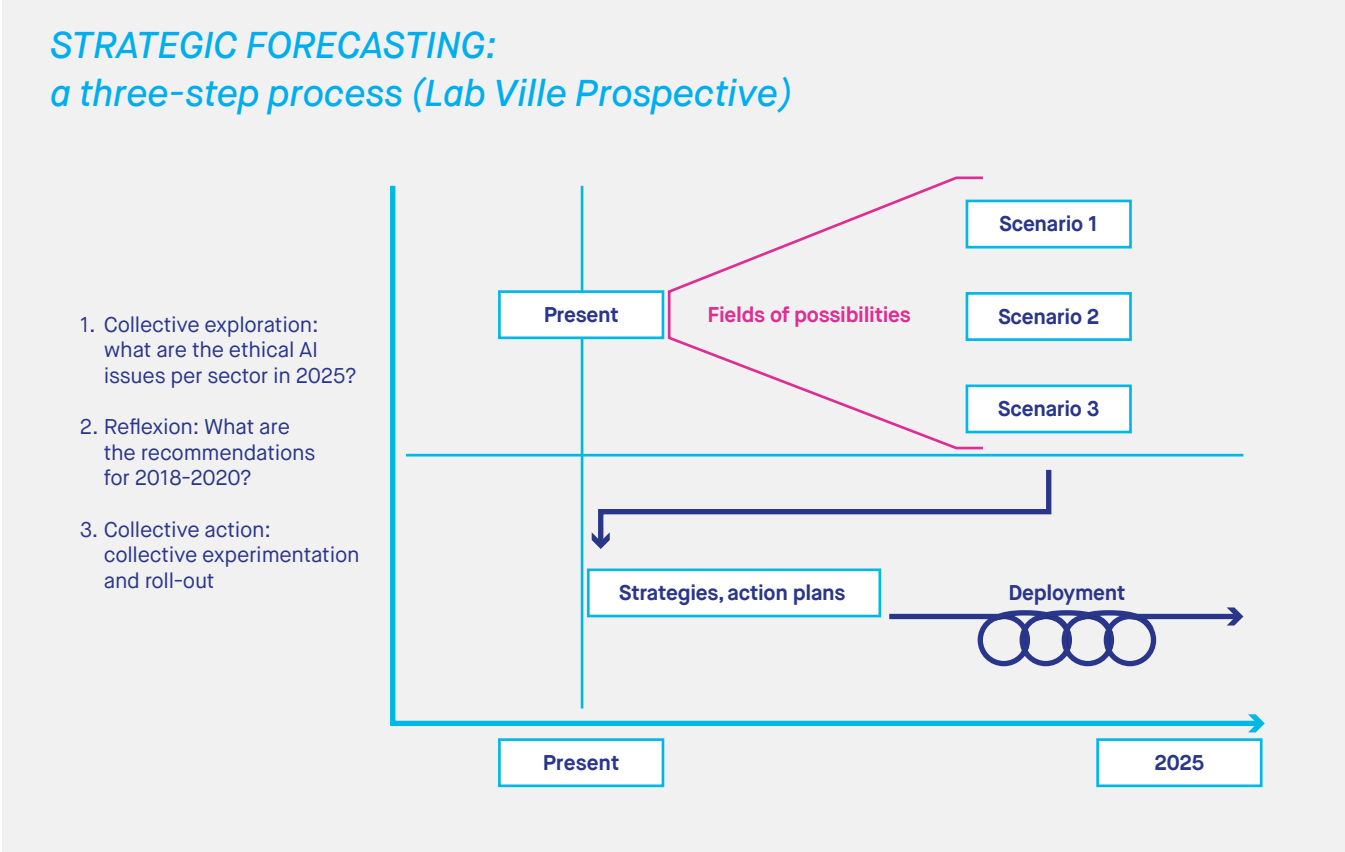
[26] Christophe Abrassart, Philippe Gauthier, Sébastien Proulx and Marie D. Martel, « Le design social : une sociologie des associations par le design? Le cas de deux démarches de codesign dans des projets de rénovation des bibliothèques de la Ville de Montréal », Lien social et Politiques, 2015, n° 73, p. 117-138.

identify, specify or anticipate ethical issues per sector on AI deployment in years to come. These 2025 discussions can then help retroactively formulate concrete recommendations for 2018–2020, to help us work toward mutually beneficial goals.

> Next, using participatory facilitation techniques in multidisciplinary "hybrid forums"[27] that invite citizens and stakeholders to reflect on shared uncertainty and possible futures (to flesh out a scenario, come up with ways to respond to an ethical risk, suggest additions to the Declaration should an orphan issue arise, i.e. without a corresponding principle).

> Lastly, paying attention to "paradigm biases" that have very powerful framing effects in how problems are viewed (e.g. tackling ethical issues of self-driving cars solely from the angle of the tramway dilemma, as MIT's Moral Machine experience team suggests) and in the context of the "speed-distance" paradigm in transport design), in order ensure a plurality of issues and draw attention to still unknown or emerging situations in a rapidly changing context.

Figure 4: Strategic forecasting: a three-step process



This goal of the co-construction workshop is to create a learning path over the course of the workshops that turns into a versatile, user-friendly and reproducible discussion kit that could be published in "open source" at the end of the co-construction process.

Details of the world cafés and co-construction days can be found in the appendix.

[27] Callon, Lacoumes, Barthe, *Agir dans un monde incertain. Essai sur la démocratie technique*, Paris, Le Seuil, 2001

4.4

WORLD CAFÉS OUTSIDE LIBRARIES

We would like to underscore the involvement of Pauline Noiseau and Xavier Boileau, two philosophy students at Université de Montréal, who have organized many world cafés in non-libraby spaces, and who used a format that encourages more organic kinds of discussions on AI issues. Moderators used very short scenarios, and hosted 2-hour sessions. These sessions sparked meaningful deliberations among citizens who wanted nothing more than to be involved in public debates, but who were rarely asked to participate in them. That's how a world café at the Maison d'Haïti, on April 25, 2018, allowed high school youth and retirees from the Saint-Michel neighbourhood in Montréal-Nord to trade opinions on AI issues. From an AI scenario on household connected objects (a smart refrigerator), this session sparked original ideas on cooking as a relational human activity, raising issues of authenticity, affection ("a touch of love") and social ability, issues that had not come up in other consultations based on the same scenario.

4.5

PORTRAIT OF PARTICIPANTS

By recruiting citizens, experts and professionals from different fields of work, we had access to a diverse pool of participants in the co-construction workshops. We were also able to contact numerous stakeholders involved in AI development through university faculties as well as inter-university research centres and their networks.

The websites and social media of our different partners played an important role in soliciting the public, although local recruitment efforts by each library involved in the project proved to be the most efficient.

Of note: nearly as many men as women took part in all workshops. The majority of participants had a post-secondary education and fell into the 19-34 age group.

Figure 5: Proportion of men and women involved in the co-construction workshops

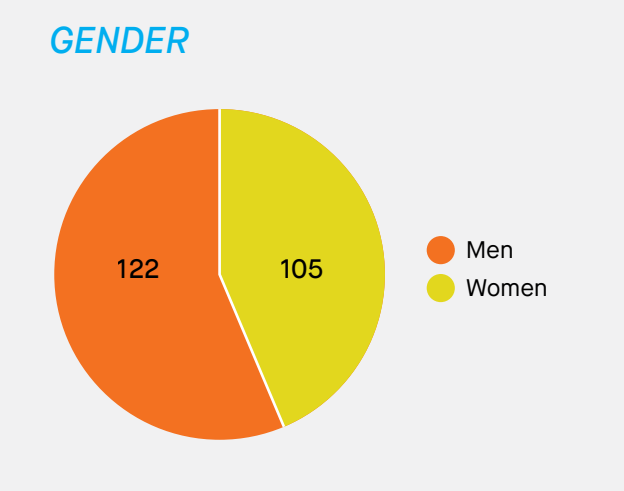


Figure 6: Participants in the co-construction workshops per age group

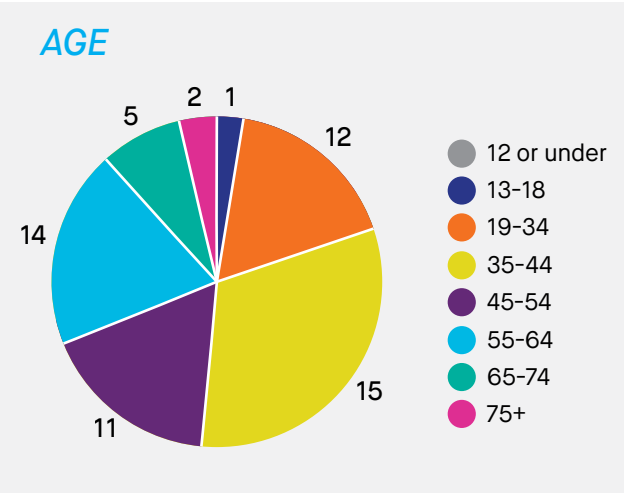


Figure 7: Distribution of participants in world cafés and co-construction days per level of education

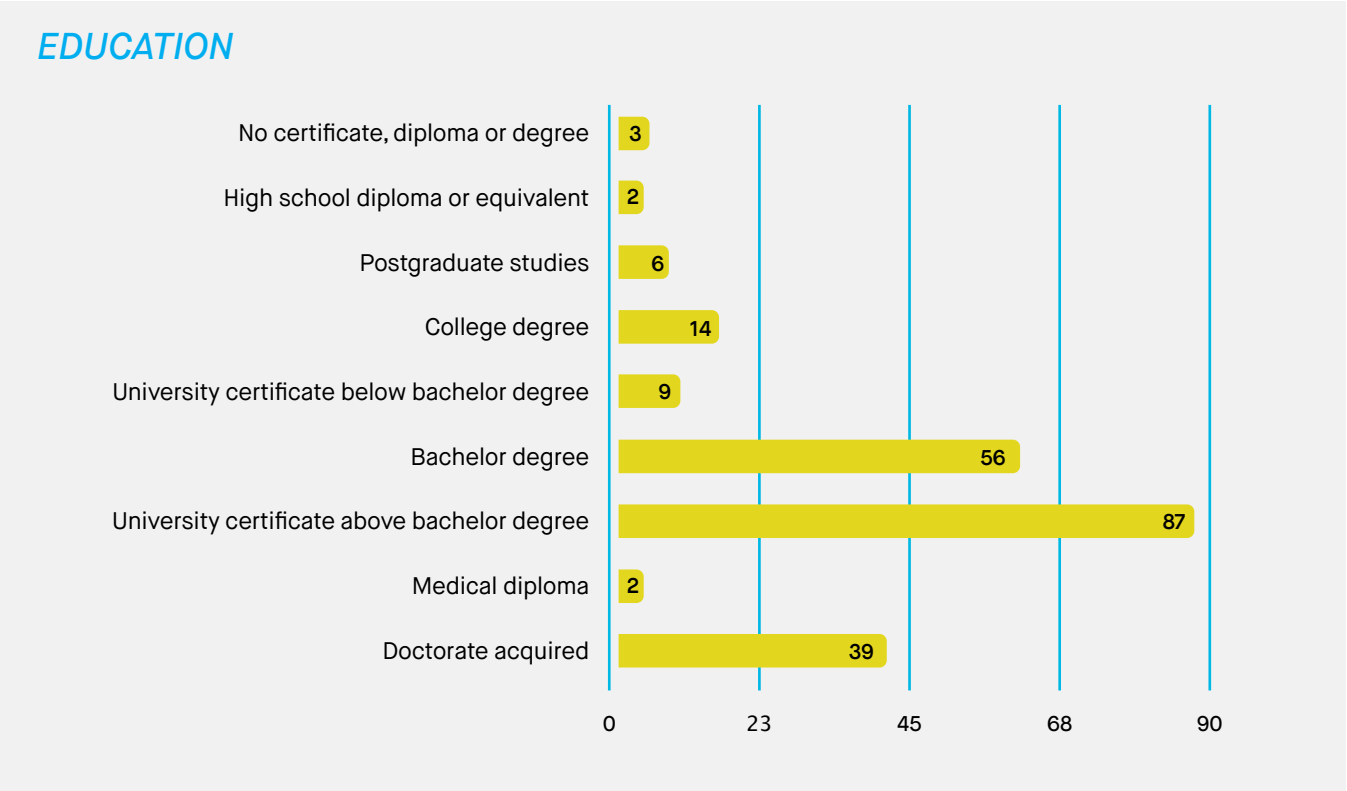
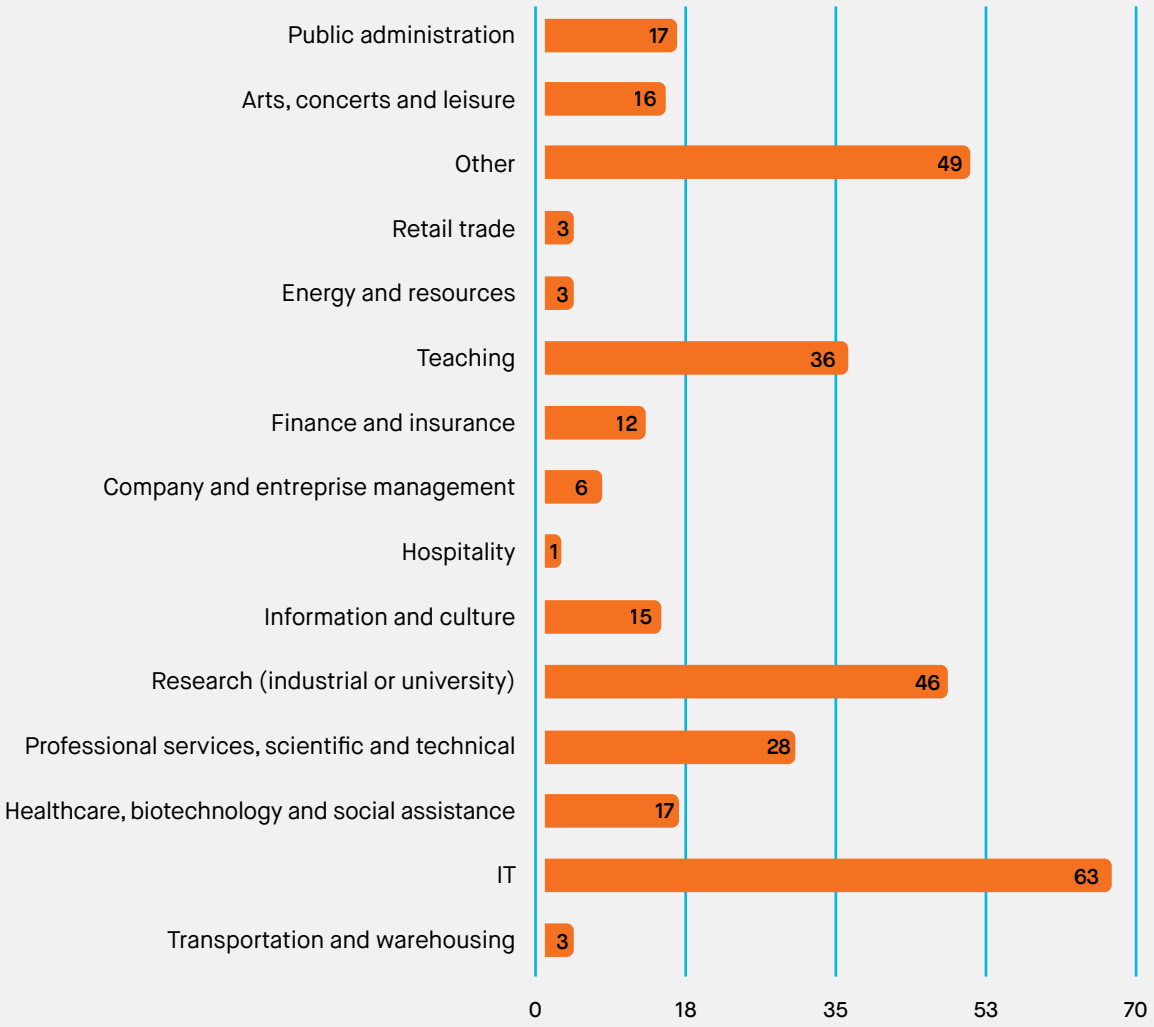


Figure 8: Distribution of participants in world cafés and co-construction days by area of activity

FIELDS OF ACTIVITY

34% of respondents indicated more than one field of activity



5. WORKSHOP DELIBERATIONS: EXAMPLES FROM SMART CITIES AND THE WORKPLACE

5.1

THE DELIBERATION PROCESS

How did the discussions and deliberations in the co-construction workshops unfold? What kinds of reactions did they elicit? What were the main points of discussion that led to recommendations for an AI framework? This section of the document includes highlights from the deliberations between participants, where each person took care to specify the reasons, principles and values justifying their position on the prospective scenario suggested as a starting point, whether it was to agree, disagree, nuance or question something. In a word, to do what pragmatic sociology has defined as justification.

To illustrate this work, the deliberations of two teams representing two of the five sectors discussed in the co-construction were selected:

a table of citizens that discussed the self-driving car (smart city sector) and a table of researchers and experts dealing with the impact of AI on jobs in businesses (workplace sector).

To formulate these recommendations, each team followed three steps where ideas were generated, then deliberated:

First step: identifying ethical and social issues per sector in 2025 (by cross-referencing the general principles of the Montréal Declaration with the 2025 user situations described in debate scenarios): determining individual issues (on Post-its), which were then expanded upon in a collective discussion where three priorities were identified.

Second step: proposing recommendations to be implemented in 2018-2020 to prepare for the responsible deployment of AI in Quebec: from formulating recommendations to choosing a few newspaper headlines.

Third step: storytelling for the deployment of the first recommendation in 2020 (the newspaper headline) to consider the "time for collective action" with its organizational constraints: from coming up with ideas to synthesizing them in an orderly fashion within a narrative.

It is important to note that between these steps and micro-steps of the deliberations, the "nature" of the ideas generated varies: some are individual intuitions (when, at the start of the exercise, participants write down many sector issues on Post-its), others stem from a collective discussion (where each person justifies their point of view) and yet others are the result of a hierarchy determined by the group (when selecting three key issues to write on the summary poster).

We note three properties of deliberative mechanisms in these foresight workshops, as discussed by Blondiaux and Sintomer in their article *L'impératif délibératif*[28] (Politix, 2002, pp. 25-26): allow new solutions to be imagined in an uncertain world; let generalities arise and aim for consensus or "deliberative disagreements" in a society characterized by value pluralism; and finally, provide a factual and normative source of legitimacy by including everyone in these deliberations.

[28] Blondiaux L. et Sintomer Y., « L'impératif délibératif », *Politix*, 2002, p. 25-26.

5.1.1 Smart city sector: self-driving cars (SDC) and sharing the road equitably

Summary of the initial 2025 scenario. In 2025, the first SDCs are circulating in Montréal and controversy arises over sharing the road and public spaces. Some lanes are now reserved for SDCs and protected by barriers, so that they can drive at a moderate, but fluid speeds (50 km/h) without the risk of accident. SDCs can also drive elsewhere, but at very slow speeds (25 km/h). Proponents of active mobility (walking, biking) disrupt these protected lanes, knowing that SDC algorithms are set to "altruistic" mode to protect people outside of them.

The goal of this scenario was to open a discussion on the ethical issues of SDCs with a situation that recreated the density and complexity of a city: low speeds and different speeds, fluidity as a priority criteria for speed, protection barriers for safety, the road as a shared space for competing uses.

The deliberations presented are the result of a three-hour roundtable in a Montréal public library, with eight citizens keen on new technologies, whose families embrace active transportation (walking, biking). From this 2025 scenario, the discussion led to an initiative presented as a headline in the March 13, 2020, edition of the Responsible AI Gazette: "First autonomous mobility literacy workshop held." What were the deliberations that led to this unique proposal? What were the defining moments? How did the ideas evolve at each step? We present and comment on certain highlights of deliberations by this team.

Highlights from first deliberations: FRAMING ETHICAL ISSUES IN 2025

A number of questions about different principles of the Montréal Declaration were written on Post-its and submitted by participants:

Autonomy:

"Will humans become too dependent when it comes to getting around?", "Will freedom of movement be impeded by AI?", "We're entrusting AI and interconnected systems with making a lot of micro-decisions to, at the expense of humans."

Well-being:

"A lot less room for spontaneity with SDCs", "How will neighbourhoods and SDC roads be developed?", "Will transportation data influence the urbanization of cities?"

Democracy and justice :

"How will building roads in working-class neighbourhoods and affluent neighbourhoods differ?", "Will only those who are well-located get to take advantage of fluid traffic?"

Privacy:

"Will we be able to track everyone's movements?" Responsibility: "Who will be held responsible for an accident?" Security: "Can fleets of vehicles be hacked?" This last principle came from participants, in addition to those found in the declaration.

Many in-depth discussions then took place, with participants bouncing off initial ideas and generating new ones on spontaneity and freedom to travel, the safety of personal data being managed by a central organization, algorithm settings and their potential for manipulation.

After a nearly 45-minute-long discussion, the participants used coloured stickers to select ethical issues for 2025, which were grouped by priority. Votes were cast via coloured Post-its on the wall and ideas associated with four principles of the Montréal Declaration were discussed, two of which were regrouped: safety, justice, and well-being and autonomy.

Table 1: Smart City, Highlights from first deliberations: framing ethical issues in 2025

2025 Ethical Issues	**1**	**2**	**3**
Description	Ease of hacking centralized system. Dilemma: collective fluidity - system vulnerability	Risk of social exclusion Settings classification by social class (e.g. trip through poor neighbourhood - VIP settings)	Loss of spontaneity when travelling, loss of independence and freedom of movement, and geo-localization.
Associated Principles	Security	Justice	Well-being and autonomy

The priority issues selected by the team are unique: although issues of security, responsibility and privacy are often raised in studies and debates on SDCs, those of justice, well-being and autonomy are not discussed as often.

Highlights from second deliberations: RECOMMENDATIONS FOR AN AI FRAMEWORK IN 2018-2020

In response to these issues, the team chose to pursue discussions by continuing to reflect on the four associated principles. Participants proposed a number of recommendations for an AI framework. We present three (out of six) here, which allow you to see how an idea evolved into a newspaper headline.

Table 2: Smart City, Highlights from second deliberations: recommendations for an AI framework in 2018-2020

Framework recommendations for 2018-2020	**1**	**2**	**3**
Description	Training for collective vigilance (e.g. driver's licence)	An all-party committee that manages incidents, infractions and other issues in democratic fashion; it must be a decision-making committee	Evaluating urban development plan during the transition period
Tool Categories	New training	New institutional stakeholder	Participative planning process

These recommendations, which demonstrate true institutional creativity (beyond the very broad examples of tools provided in the participant booklet), dovetail with issues identified in the previous step, but also generated more robust ideas (they are not merely deductions from tools adapted from an identified ethical case). The idea of vigilance training and participating in collective decision-making (through an all party committee and open planning) led to recommendations for capacity-building and local forms of democracy.

Highlights from third deliberations: WRITING A HEADLINE AND LEAD FOR A 2020 NEWSPAPER

These measures were then storyboarded on the poster. The headline of the March 13, 2020, edition of the Responsible AI Gazette designed by the team read as follows:

"FIRST AUTONOMOUS MOBILITY LITERACY WORKSHOP HELD"

"The Quebec public library network has introduced a training program on the use of self-driving cars. On the curriculum: collective vigilance; the code of ethics; how to get involved in the city's decision-making process; roadsharing between pedestrians, bicycles, SDCs, trucks; understanding the rules; trial sessions; the issue of algorithm settings."

This newspaper headline, which was drafted following a discussion among the participants, once again demonstrates how ideas evolve. The principle of a workshop on **"autonomous mobility literacy" allows new meaning to be created** by integrating various recommendations proposed in the previous step, thus widening the scope to autonomous mobility and not merely SDCs (thereby including the possibility of autonomous multimodal transportation). This headline also presents a **collective action measure** with a progress target (training and capabilities of citizens, the possibility of joining city decision-making committees on SDC deployment) and organization (roll-out in public libraries across Quebec, which are now transforming into cultural "third spaces" for citizens).

The result of this roundtable is particularly interesting because it allows us to consider the ethical question of self-driving vehicles from the perspective of autonomy and social justice in the city, and not strictly responsibility in the event of an accident, as MIT's Moral Machine initiative does, for example, with the moral dilemma of the tramway[29].

[29] MIT site: http://moralmachine.mit.edu

5.1.2 Workplace sector: Socially responsible restructuring?

Summary of the initial 2025 scenario.
In 2025, many businesses use AI in their management tools. Such is the case for an eco-friendly logistics company that must make a massive investment in AI and robotics to remain competitive. Parcel sorting, routing, administrative follow-up, calculating the carbon footprint of the trips, self-driving electric trucks: in total, up to one third of the company's positions could be cut. The company, which is very socially involved, wants to proceed with restructuring in a socially responsible manner, for instance by creating a data processing co-op to rehire as many salaried employees as possible, independently from the big corporations. Will it be able to do so in time?

The goal of this scenario was to spark a discussion on the ethical and social challenges related to change processes caused by AI that thousands of SMEs and big businesses in Quebec will face between 2020-2030.

The deliberations presented in this section were held during a day-long roundtable in Montréal that brought together nearly 10 researchers and experts working on workplace mutations, social participation and social responsibility of businesses and unions. A citizen that had previously attended a workshop in a public library also took part in this roundtable.

Working from this 2025 scenario, this team came up with an initiative that made the headline of the February 18, 2020, Responsible AI Gazette: **"First measures of the joint interdepartmental committee on responsible digital transition."** As in the previous case on self-driving cars, how did the deliberations lead to this unique proposal? What were the defining moments? How did the ideas evolve at each step? We present and comment on highlights of the team deliberations.

Highlights from first deliberations: FRAMING ETHICAL ISSUES IN 2025

Participants wrote numerous ideas on Post-its in the first half of the morning workshop. Here are a few of them and a sampling of some ideas taken from Post-its, which were grouped according to principles of the Montréal Declaration.

Some ideas were associated with different principles of the Montréal Declaration :

WELL-BEING:

"What should we promote? The company or society?", "Adopting different perspectives on well-being: individual (employee), social and collective development, economic development (SME)", "What does performance measurement look like when robots or co-bots never get tired, unlike humans?", "What are the possible positive aspects: professional support, e.g. in medicine, less drudgery in certain positions", "What are the new forms of work and protection with work/leisure?"

AUTONOMY:

"What sort of career and life paths? Can you choose not to change careers because of AI? What are the consequences?", "collective autonomy: for collectively and critically anticipating discussion on the urgency of adapting"

RESPONSIBILITY :

"Who is held responsible for these changes?", "Is the social and ethical responsibility of the transition individual—each company–or collective—society, the government?", "Where does for the transition funding come from?"; "How can we align cost-effective measures and responsibility in an emergency situation?"

KNOWLEDGE:

"What will collaboration between humans and robots look like? Workload, health and safety, training, acceptability, cybersecurity," "How is data collected in settings where this type of work is mainly carried out by private corporations (GAFAM)?", "How can we prevent people from getting pigeonholed in classes?", "What are the possibilities of data being shared?", "What is the impact on the educational system?"

THE JUSTICE PRINCIPLE:

"What independence exists when power is concentrated among GAFAM?", "How will the benefits of AI be distributed among society?", "Will productivity gains created by AI and industry 4.0 be sufficient to fund social change if companies engage in tax evasion?", "What fairness exists when sharing and coding an employee's implicit knowledge so that it can be transform into data or feed automation?", "Do we have a choice, as employees, not to reveal this information?", "What is the criteria for choosing those who are replaced and those who are trained?", "Will social protection of tomorrow be accessible?", "What access to rights, such as right of association, in this new, reorganized workplace?"

DEMOCRACY:

"Is job insecurity inevitable when transition can be anticipated?", "the politicized short-term vision over a long-term vision", "obscuring decision-making processes", "risks of bias in algorithm training sets", "the need for a democratic debate"

We should note that the classification of the Montréal Declaration on responsible AI principles was useful in providing discussion benchmarks, and that the participants even came up with unique issues on certain principles: the need to address well-being and responsibility for the transition from different points of view (individual and collective); the relationship with social time, with collective anticipation and the opaque language of urgency opposed to one another, as a condition of our collective autonomy and exercising our democracy (lack of time preventing well-informed democratic work); a strong need for justice in the social redistribution of AI benefits, namely in terms of equity accompanying codification, and therefore possible automation, of employee skill sets.

After a good hour of discussion, participants used coloured stickers to determine groups of 2025 ethical issues they deemed priorities. With votes spread out fairly evenly on various issues, all deemed equally important by the group, three priorities for the poster were identified after the ideas discussed in the first half of the workshop were synthesized (see table below).

Table 3: Workplace, Highlights from first deliberations: framing ethical issues in 2025

2025 Ethical Issues	**1**	**2**	**3**
Description	**Massive concentration of power** (see GAFAM) that prevents: - Equitable sharing of AI benefits - Arrival of new stakeholders (new co-op type business models) - Inequities to be minimized (literacy)	**Technological determinism, inevitability ("Black box society") and urgency:** instead of taking the time to have an informed, participative, democratic debate on new social risks, social development models, performance measures, work experience.	**Defining the common good and the kind of collective responsibility in the digital transition** For example: Which stakeholders? The company alone? The State? Unions? The educational system?
Associated Principles	Justice and independence	Democracy, knowledge and collective autonomy	Well-being and responsibility

Second deliberations: RECOMMENDATIONS FOR AN AI FRAMEWORK IN 2018-2020

To respond to these issues, the team resumed talks in the afternoon by leading another roundtable during which participants drafted recommendations for an AI framework. This then led to numerous recommendations that were discussed one by one as a group. The table below presents an excerpt (six out of the more than 10 proposals that were formulated by the group), in order to follow the evolution of an idea up to a newspaper headline.

Table 4: Workplace, Second deliberations: recommendations for an AI framework in 2018-2020

2018-2020 framework recom-mendations	**1**	**2**	**3**	**4**	**5**	**6**
Description	**Reinforce digital literacy for all.** With a skill set reference software for public libraries, schools, and the workplace. By tackling the question of illiteracy and "non-use" by citizens.	**Joint permanent interde-partmental committee on AI, executive next to PM.** Interface between themes of economy, employment, education and culture. (see digital strategy)	**Digital AI insurance funds to make way for training and transition.** Type of measure: a 50-week Parental Insurance Plan which can also generate a minimum income to prevent job insecurity.	**Incentives on new business models for data processing.** Example: Co-op model that prevent self-employed workers processing data from becoming isolated and guarantees collective autonomy.	**Investing in responsible AI for the common good.** SRI (Socially responsible investment) model. Investments from the State, individuals, in synergy with the worker's fund.	**Accelerated process to update and create professional programs.** With cégeps, universities, departments, professional orders impacted by AI (e.g. law, healthcare).
Tool Categories	New training	New institutional stakeholder	New insurance mechanism	Incentive	Funding device	Planning process

As in the previous case of self-driving cars, these recommendations, which demonstrate true institutional creativity (beyond the very broad examples of tools provided in the participant booklet), dovetail with issues identified in the previous step, but also generate more robust ideas. If digital literacy is indeed a goal in the policy's agenda (e.g. Stratégie numérique du Québec), the need to expand it was highlighted. Other recommended measures are innovative and call for the creation of new public, all-party or collective measures to ensure the true autonomy of Quebec society on AI issues in the workplace. To that end, the group chose collective responsibility towards AI in its transition into society.

Third deliberations: WRITING A HEADLINE AND LEAD FOR A 2020 NEWSPAPER

These measures were then storyboarded for the poster. The headline drafted by the team for the February 18, 2020, Responsible AI Gazette reads as follows:

FIRST MEASURES OF THE JOINT INTERDEPARTMENTAL COMMITTEE ON RESPONSIBLE DIGITAL TRANSITION

The new committee, created on March 14, 2018, after the co-construction workshop for the Montréal Declaration for Responsible AI, quickly got to work and developed a coherent strategy integrating all stakeholders. In early 2020, the committee was proud to announce the launch of four (4) programs:

1. A new digital insurance fund worth $2 billion (funded by productivity gains attributed to AI).
2. An agreement with all cégeps and universities to accelerate the renewal of training programs.
3. A support program to create self-employed worker co-operatives (against job insecurity).
4. A five-year literacy fund worth $10 billion based on a new skill set inventory.

This newspaper headline, which was drafted following discussion among participants, once again helps ideas evolve. The joint interdepartmental committee on responsible digital transition was entirely new. This new institutional stakeholder, born of a reflection on a 2025 scenario that dealt with the impact of AI on the Quebec workplace, could represent a new common step for many public policies that successfully address digital transition and digital literacy, but do not tackle the question of AI's social impact: the **Stratégie numérique du Québec** du ministère de l'Économie, de la Science et de l'Innovation (MESI), the **Stratégie nationale sur la main-d'œuvre 2018-2023** du ministère du Travail, de l'Emploi et de la Solidarité sociale (MTESS), the **Plan stratégique 2017-2022** du ministère de l'Éducation et de l'Enseignement supérieur (MEES). This new stakeholder, possibly the result of a cross collaboration between the Commission des partenaires du marché du travail (CPMT), the Comité consultatif sur le numérique and the Commission mixte de l'enseignement supérieur, would specifically anticipate workplace transformations and new training and adaptation issues caused by the deployment of AI in Quebec's public and private organizations.

6. PARTICIPANTS IN THE CO-CONSTRUCTION AND WORKING GROUPS

Citizens, professionals and experts who took part in the workshops, in Québec and Paris, and agreed to let us publish their names:

Sihem Neila Abtroon

Sébastien Adam

Béatrice Alain

Hassane Alami

Rana Alvabi

Alejandro Arreola-Alvarado

Gabriel Arruda

Jean-Claude Asssaker

Barthélémy Aucourt

Naomi Ayotte

Manon Babine

Maryluisa Barillas

Philippe Beauchemin

Stéphane Beaulieu

François Beauregard

Claude Bédard

Sylvain Bédard

Abdelkader Bekhti

Halim Benzaïd

Vincent Bergeron

Alexandre Berkesse

Karl Bherer

Emmanuel Bloch

Marise Bonenfant

Serge Bouchard

Caroline Boudreault

Lyne Bourbonnais

Véronique Boutier

Morgane Bravo

Robert Bruno

Beatrice Cassar

Ofelia Castaneda

Chantal Caux

Christian Chabot

Michel Chabot

Karine Charbonneau

François Charbonnier

Anne Chartier

Philippe Chartier

Guillaume Chicoisne

Pierre Choffet

Dominic Cliche

Lilen Colombino

Cristina Cotargasanu

François Côté

Jacques Coulombe

Lise Couturier

Alexis Cuglietta

Christian Cyr

Yvonne Da Silveira

Geneviève Dagneau

Hélène David

François-Michel De Rainville

David Décary-Hétu

Guillaume Déraps

Yves B. Desfossés

Michel Desy

Marc-Antoine Dilhac

Maxime Duban

Jean-Yves Dubé

Geneviève Dubois-Flynn

Mathieu Dubreuil-Cousineau

Geneviève Dufour

Arnaud Duhoux

Annie Dulude

Laurence Dumont

Mathieu Dumouchel

Benoit Dupont

Nicolas Dupras

Diane Duquette

Irina Entin

Julian Falardeau

Jacqueline Forien

Simon Frappier

Benoit Gagnon

Marie-Pierre Gagnon

Marina Gallet

Hortense Gallois

Sébastien Gambs

Véronique Gareau-Chiasson

Mathieu Gauthier-Pilote

Sylvie Gélinas

Thomas George

Gueno Gianni

Jean-François Gignac

Martin Gibert

Patricia Gingras

Béatrice Godard

Christian Goudreau

Gilles Gouin

Mervine Gowry

Alexandre Gravel

Michel Grou

Alexandre Guédon

Pascaline Guenou

Pierre Guillou

François Guité

Carl Hamilton

Simon-Pierre Harvey

Lucie Hébert

Ghiles Helli

Lucas Hubert

Aida Issa

Sabrina Jocelyn

Erwan Jonchères

Nico Julien

Debbie Jussome

Ed Khazen

Amy Khoury

Frederic Kleindienst

Andrée Labrie

Anne-Marie Lacombe

Marie-Claude Lagacé

Henri Lajeunesse

Karine Landry

Jean-Michel Lapointe

Jonathan Lasprilla

Sylvie Lavoie

Jean Latière

Louis Lecaer

Dominique Leclerc

Sarah Legendre Bilodeau

Pascale Lehoux

Claude Lejeune

Mélanie Levasseur

Elisabeth Limoges

Pamela Lirio

Robert Locas

Santiago Lopez

Aurélie Macé

Aicha Mafhoum

Suzanne Mainville

Mantas Manovas

Mathieu Marcotte

Jean-Pierre Marquis

Cloderic Mars

Marie Martel

Marième Mauger-Lavigne

Moussa Mekhnach

Natacha Mercure

Bruno Milia

Michael David Miller

Ann Mitchell

Erica Monteferrante

Farida Mostefaoui

Maria Moudfir

Jocelyne Mouton

Khalil Mouzawak

Vanessa Murray

Orly Nahmias

Vanessa Nantel

John Newhouse

Justin Ngoza

Zoonie Nguyen

Lisa Marlène Ntibayindusha

Catherine Olivier

Daniel Pascot

Florence Paulhiac

Ludovic Penet

Jorge Perez

Caroline Pernelle

Lorenzo Perozzi

Geneviève Perreault

Benoit Petit

Emmanuel Picavet

Louis Piette

Frédérick Plamondon

Pier-Luc Plante

Kamila Podgorska-Gilbert

Keith Poitras

Julie Politi

Philippe Polveche

Thomas Poulin

Denis Poussart

Emmanuelle Praine

Louis-Philippe Pratte

Mariel Ramos

Diane Raymond

Catherine Régis

Laurence Renault

Cassie Rhéaume

Toussaint Riendeau

Anne-Marie Robert

François-Xavier Robert

Louis-Nicolas Robert

Nicolas Roby

Stéphane Roche

Marie Roy

Sara Russo-Garrido

Laurence Sabourin

Iger Sadoune

Marie-Noëlle Saint-Pierre

James Sangster

Sylvie Saucier

Anton Selikhov

Jean-François Sénéchal

Eric Shannon

Danielle Sicotte

Chantale Simard

Julie Simard

Jean-Hébert Smith-Lacroix

Karima Smouk

Yanis Taleb

Isabelle Tanba

Christian Tanguay

Marc Tomkinson

Daniel Tremblay

Jérémy Trudel

Marie-Christiane Trudel

Félix Vaillancourt

Julie Verdy

Arnaud Vicari

Danael Villeneuve

Grant Wark

Bryn Williams-Jones

Lemy Wong

William Wong

Almina Yagoubi

Ming Yue

Co-construction team – In Québec and Paris

Simon Beaudoin-Gagnon, Maison des étudiants canadiens

Alexandre Beaudoin-Peña, Université de Montréal

Bhavish Beejan, Université Laval

Liam Bekirsky, Maison des étudiants canadiens

Karl Bherer, Université Laval

Alexis Bibeau, Université Laval

Pierre-Antoine Boutin-Panneton, Université Laval

Katie Charpentier-Bourque, Maison des étudiants canadiens

Arnaud Brubacher-Chouinard, Maison des étudiants canadiens

Dominic Cliche, Université Laval

Valentine Crosset, Université de Montréal

Rosemarie Desmarais, Maison des étudiants canadiens

Eve Gaumond, Université Laval

Martin Gibert, IVADO and Centre de recherche en éthique éthicien

Emilie Guiraud, Université Laval

Haykuhi Gutrez, Maison des étudiants canadien

Hubert Hamel-Lapointe, Université de Montréal

Audrey Houle, Université Laval

Samira Illourman, Maison des étudiants canadiens

Nico Julien, Université Laval

Henri Lajeunesse, Université Laval

Lauriane Long-Raymond, Maison des étudiants canadiens

Guillaume Macaux, Université Laval

Vincent Mai, Université de Montréal

Marième Mauger-Lavigne, Université de Montréal

Christophe Mondin, CIRANO

Orly Nahmias, citizen

Pauline Noiseau, Université de Montréal

Judith Paquet, Université Laval

Pierre-Luc Plante, Université Laval

Léa Ricard, Université de Montréal

Lynda Robitaille, Centre de recherche en données massives, Université Laval

Jason Stanley, Université de Montréal

Yanis Taleb, Université de Montréal

Clémence Varin, Université Laval

Nathalie Voarino, Université de Montréal

Camille Vézy, Université de Montréal

Alessia Zarzani, Université de Montréal

Consulted experts

Sylvain Bédard, Patient Coordinator at Centre of Excellence on Partnership with Patients and the Public (CEPPP)

Louise Béliveau, Vice-Rector of Student and Academic Affairs, Université de Montréal

Guillaume Chicoisne, Scientific Programs Director, IVADO

David Décary-Hétu, Assistant Professor, School of Criminology, Université de Montréal; Regular researcher at Centre international de criminologie comparée

Pierre-Luc Déziel, Professor, Faculty of Law, Université Laval

Thierry Karsenti, Full Professor, Faculty of Education, Université de Montréal; Chairholder of the Canada Research Chair on information and communication technologies in education

Jihane Lamouri, Diversity Coordinator, IVADO

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

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

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

Nicolas Merveille, Professor in École des sciences de la gestion (ESG), Université du Québec à Montréal (UQAM); Co-chairholder of the International Life Cycle Chair, ESG UQAM and École Polytechnique; Supervisor for the City of Montréal's Internet of Things Ethics and Social Acceptability profect

Gregor Murray, Director of Industrial Relations and Interuniversity Research Centre on Globalization and Work (CRIMT)

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

Nicolas Roby, Scientific Coordinator of Industrial Relations and Interuniversity Research Centre on Globalization and Work (CRIMT)

Frank Scherrer, Full Professor, Urbanisme, Université de Montréal; Director at the École d'urbanisme et d'architecture de paysage, Université de Montréal; Academic Director of the EDDEC Institute, an organization promoting the environment, sustainable development, and circular economies

Marie-Odette St-Hilaire, Architecte de solutions TI, Science de données, Service des technologies de l'information, Ville de Montréal

Management team

Isabelle Bayard, Vice-Rector assistant of Reasearch, Discovery, Creation and Innovation, Université de Montréal

Joliane Grandmont-Benoit, Project Coordinator, Vice-rectorate of Student and Academic Affairs, Université de Montréal

Anne-Marie Savoie, Adviser, Vice-Rectorate of Reasearch, Discovery, Creation and Innovation, Université de Montréal

Research and analysis team

Valentine Crosset, Doctoral student in criminology, Université de Montréal

Jean-François Gagné, Programs Adviser, Political Science, Université de Montréal

Vincent Mai, Doctoral student in robotics, Université de Montréal

Mario Ionut Marosan, Master in political philosophy, Université de Montréal

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

Loubna Mekki-Berrada, Doctoral student in neuropsychology, Université de Montréal

Christophe Mondin, Research professional for CIRANO

Camille Vézy, Doctoral student in communication, Université de Montréal

Nathalie Voarino, Scientific Coordinator, PhD Candidate in Bioethics of Université de Montréal

Alessia Zarzani, Doctoral student in Planning, Université de Montréal

Coordination team in Paris

Jacques-Henri Gagnon, Head, Communication, Youth and Academic Relations, Embassy of Canada to France

Hanane Hadjiloum, Communications Officer, Maison des étudiants canadiens

Christine Métayer, Director of the Maison des étudiants canadiens

Clément Thiébault, Trade Commissioner, France (ICT), Embassy of Canada to France

Partners who contributed to the fall co-construction workshops

Students from the Comité intersectoriel étudiant (CIÉ) from the Québec Research Funds, participants in the Journées de la relève en recherche held by ACFAS

Professional members from the Coalition for the Diversity of Cultural Expressions (CDCE-Canada)

Elected officials and employees from trade unions that participated in the AI discussion forum organized by the Syndicat de la fonction publique et parapublique du Québec (SFPQ)

ANNEXES

ANNEX 1 – CO-CONSTRUCTION WORKSHOPS: HOW THEY WORK

World cafés

World cafés are three-hour-long meetings in public libraries. These meetings are inclusive, open to all citizens, and take place in a friendly atmosphere. These meetings will be based on the World Café model.

The world café provides an enjoyable forum for conversation and seeks to encourage constructive dialogue and the exchange of ideas. The goal is to recreate a café ambiance where participants debate a question in small groups. At regular intervals, participants change tables. One host stays at the table and sums up the previous conversation for the people who have just arrived. The ongoing conversations are therefore "pollinated" by ideas from previous conversations. At the end of the process, the main ideas are summed up during a plenary session, and possible follow-ups are submitted for discussion[30].

This world café technique was adapted and enhanced by many elements:

- Presenting the Montréal Declaration and the social and ethical issues of AI
- Reading sector foresight scenarios set in 2025 to kick off discussions
- Using a poster to document the discussions
- Handing out a participant workbook presenting the principles of the Montréal Declaration for Resonsible AI, a lexicon and an sample classification of possible recommendations.

Here is what a typical world café looks like:

Tableau 5 : Déroulement type des cafés citoyens

Steps	Time	Description
Welcome	1 p.m. to 1:30 p.m.	Coffee and snacks
Discovering AI and its ethical and social implications	1:30 p.m. to 2 p.m.	**Educational Introduction:** introduction to the ethical and social implications of artificial intelligence (Montréal Declaration), presentation of scenarios set in 2025 and the activity.
World café	2 p.m. to 4 p.m.	> Four thematic islands (on AI in health, justice, education, smart cities and the workplace) are hosted by a facilitator. Each island hosts a small group of participants (6 to 10) for two 50-minute discussions about an AI scenario set in 2025. > participants are invited to imagine the "front page of a 2020 newspaper" (headline and first paragraph) discussing an important initiative in Quebec for the responsible deployment of AI.
Summary in plenary session	4 p.m. to 4:30 p.m.	**Summary of discussions during a plenary session.** The facilitators sum up the posters from each thematic island, followed by a group discussion.

[30] Definition from the Institut du nouveau monde (INM)

Co-construction workshops

These one-day meetings brought together citizens, stakeholders and experts seeking to explore sector issues and develop recommendations. They are based on the foresight co-design model, developed at the University of Montréal's Lab Ville Prospective.

The workshops are based on the foresight co-design model that combines design, participation and forecasting: imagining scenarios and unknown prototypes as conversation starters, opening up cognitive possibilities and pathways for exploration (the design dimension); using collective participation techniques that bring together stakeholders from diverse backgrounds, citizens and organizations as experts (for the collective aspect of the "co"); finally, the foresight approach which consists of projecting oneself into a possible future 10 or 20 years from now to make an imaginary detour and then work back from there to develop innovative paths that link the present to the most desirable futures. Michel De Certeau, in his work La culture au pluriel (1993, p. 223) highlights otherness in foresight, saying, "the future engages the present in the mode of otherness". And Georges Amar, in an article on conceptive foresight (in Futuribles, 2015, p. 21) insists on the importance of creating a narrative around the unknown to build an open future: "We prefer inefficient known properties to the promising unknown. The purpose of foresight is to work on the unknown, to give words, concepts, language on it. So that while it remains unknown, it becomes more accessible, leads to reflection, and action.'

Here is what a typical co-construction workshop looks like:

Table 6: What a typical co-construction workshop looks like

Steps	Time	Description
Welcome	8:30 to 9 a.m.	Coffee and pastries
Introduction and AI Discovery	9 a.m. to 10 a.m.	**Introductions:** principles of artificial intelligence, ethical issues surrounding AI (Montréal Declaration) and foresight scenarios.
Foresight team	10 am to 11:30 am	**Team foresight:** starting with a trigger event and the *Montréal Declaration* principles, frame the ethical and social issues raised by the 2025 scenario and explore how an ethical controversy could arise or grow.
	11:30 am to 12:30 pm	**Plenary:** Plenary presentation of 2025 ethical and social issues raised, and discussions with entire group.
Lunch on site	12:30 pm to 1 pm	Lunch
Developing recommendations	1:30 pm to 2:45 pm	**Developing recommendations** Work in teams: using the 2025 ethical issues identified in the morning, develop recommendations (rules, sectoral codes, labels, public policies, research programs, etc.) to implement in 2018–2020 in Quebec.
	3 pm to 4 pm	**Plenary team presentations** and group discussion
Conclusion and follow-up	4 p.m. to 4:30 p.m.	Review and observations surrounding the day

ANNEX 2 –
FORESIGHT SCENARIOS: WINTER CO-CONSTRUCTION WORKSHOPS

SCENARIO WRITING TEAM

Christophe Abrassart, Scientific Co-director of the Declaration, 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)

Valentine Crosset, Doctoral student in criminology, Université de Montréal

Marc-Antoine Dilhac, Scientific Co-director of the Declaration, Full Professor, Department of Philosophy, UdeM, Chair of the Ethics and Politics Group, Centre de recherche en éthique (CRÉ), Canada Research Chair in Public Ethics and Political Theory

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

Vincent Mai, Doctoral student in robotics, Université de Montréal

Christophe Mondin, Research professional for CIRANO

Nathalie Voarino, Scientific Coordinator, PhD Candidate in Bioethics of Université de Montréal

Camille Vézy, Doctoral student in communication, Université de Montréal

Alessia Zarzani, Doctoral student in Planning, Université de Montréal, and PhD in Landscape and Environment, Université la Sapienza de Roma

This annex presents a summary of all the AI scenarios used in this first co-construction phase, and five complete scenarios. Set in 2025, in Quebec, they were the starting point for debates and deliberations on the ethical questions raised by artificial intelligence. The year 2025 was selected as it was in the near future, at the heart of the decade 2020–2030 which should see intensive deployment of artificial intelligence in society.

1. Scenarios per theme

Eighteen scenarios were debated from February to May 2018. The table below presents a summary of these scenarios.

Table 7: Scenario summaries

Theme	2025 AI scenario	Summary of AI scenario in 2025 in Quebec
1. Predictive Health	Healthy digital twins	Olivier learns that one of his 126 digital twins has been diagnosed for depression. Should he go see a professional?
	Discriminating Health Insurance	Olivier's insurance company asks him to change his lifestyle, based on his personal data. Can he refuse without suffering any consequences?
	Vigilo, a house robot for the elderly	Soline is 80 years old and lives at home with Vigilo, her robot companion. Her robot regularly reports predictive diagnoses on Soline's health to her family. Does she want everything revealed?
	A therapeutic decision at the hospital	An experienced doctor and a medical recognition algorithm do not quite agree on a diagnosis.
2. Smart City	Self-driving cars (setting the algorithm and sharing the road)	To ensure its zero-accident policy, the city has established safety barriers on roads where self-driving vehicles can go "fast" (50 km/h). A controversy on sharing the road ensues.
	Self-driving cars (restricted use)	Self-driving cars have become a rideshare service for citizens. Priority access criteria is managed by AI to maximize the city's predictive economic growth.
	A connected fridge that wants what's best for you (nudges)	A family purchased a smart fridge with a "nudge" program to encourage healthy eating and reduce risks of disease. How will the gains from this system be divided between the insurance company and the family?
	A social rating based on a carbon footprint	A family's consumption is defined and tracked to prevent a negative impact on the environment.
	A smart toy that's not all that loyal!	How far does a smart toy's loyalty to a child go? Is it the same as a friend's?

Theme	2025 AI scenario	Summary of AI scenario in 2025 in Quebec
3. Predictive education	AlterEgo, AI that assists learning at school	AI helps students learn more efficiently, thanks to personalized homework and exercises. Does the teacher still have complete professional autonomy?
	AlterEgo2, AI School Guidance Assistant	AI guides students towards careers where the odds of succeeding are very strong. Based on their history of school data, will the choice really reflect the student's wishes?
	Nao, AI that helps prepare conferences	AI helps a lecturer develop his presentation and update it throughout the lecture, according to the reactions of his students.
4. Police and predictive justice	A preventive arrest in a public space	Cross-referencing Alexandre's personal data has recently flagged him as an individual who is potentially at risk. After acting strangely in a public space, he is arrested preventively.
	A parole decision	A judge makes the decision to order probation for a detainee, against the algorithm's recommendation. The algorithm anticipates likely recidivism, but without taking into consideration a new reinsertion program (without any data history).
5. Workplace	AI to improve workplace atmosphere	A company's human resources department uses AI with data mining to evaluate the behavioural style of their employees and help them to cultivate a "good workplace atmosphere".
	Recruitment AI as a compulsory step to employment	All candidates for a position will be recruited according to a video analyzed by AI, in order to eliminate any bias, favourable or not. Is recruitment neutrality real, and is it desirable?
	Socially responsible restructuring	A sustainable logistics company must massively incorporate AI into many of its services to remain competitive. But it wishes to do so in socially responsible fashion.
	A new committee on professional development	A company's professional development committee welcomes new members: the representatives of collaborating robots. Not everyone shares the same opinion about this change.

2. Five full scenarios

The five scenarios selected each explore a possible situation in 2025 for one of the themes discussed in the first co-construction phase of the Montréal Declaration: predictive health, predictive education, smart city, predictive justice, and the transversal theme of changes in the workplace.

Each scenario presents the story of a case that was built by integrating numerous dimensions: a sector problem, a user experience set in 2025, a learning apparatus that uses data mobilization and one or more artificial intelligence techniques, and finally, ethical and social issues.

Table 8: Elements of the five scenarios

2025 AI Scenarios	Digital twins	Self-driving cars	AlterEgo	Parole	Responsible restructuring
Themes	1. Predictive healthcare	2. Smart City	3. Predictive education	4. Predictive justice	5. Workplace
Sectoral issue	Preventive and personalized healthcare using similar profiles	Safety and sharing the road	Personalized learning at school	A judge's decision in the case of uncertainty	Preventive and socially responsible management of changes
Types of AI learning	Clustering data into homogenous groups through unsupervised learning	Algorithms of self-driving cars for vision, decision-making supervised and reinforced (learning)	Supervised teaching (student concentration) and reinforcement (homework follow-up policy)	Supervised teaching of past cases of recidivism	All AI from the moment it creates changes in companies and administrations
Ethical and social issues (examples)	Privacy: data confidentiality	Justice: equitable sharing of public spaces	Privacy: confidentiality of student data	Autonomy and critical knowledge in decision-making	Justice: equitable sharing of productivity gains

Theme 1:
PREDICTIVE HEALTH

Initial scenario:
DIGITAL TWINS

MARCH 10, 2025. Olivier receives a notification on his phone that one of his digital twins has just been diagnosed with depression. Digital twins are people who share the same biological traits and have similar health profiles. All data pertaining to Oliver's health has been collected by Health Canada since December 2023. Some is provided by his phone's health app (such as the number of steps taken in a day, or the number of hours of sleep), and from what he shares publicly on social media (data purchased from Alphabet and Baidu). They are cross-referenced with data provided directly from the healthcare system regarding his disease history and genetic predisposition. This data is linked to that of the entire population in the "world health cloud", overseen by the World Health Organization since 2023, which helps define individual health profiles to offer each person targeted and highly personalized prevention and precision medicine.

Olivier thus discovers that morning that he is at risk of developing the same pathology as one of his 126 digital twins. Faced with this prognosis, Health Canada's algorithm recommends that Olivier go to a mental health clinic to receive a personalized preventive treatment, reduce his workload to less than 40 hours a week, and increase his physical activity, given the proven beneficial effects of sports to prevent depression. Olivier decides to ignore this advice, as he is working on a contract that could have major repercussions on his career. However, over the course of that week, he learns that 25 of his digital twins have received a similar diagnosis.

Theme 2:
SMART CITY

Initial scenario:
SELF-DRIVING CAR – SETTING THE ALGORITHM AND SHARING THE ROAD

FALL 2025. LThe Plateau-Mont-Royal and the Rosemont—La Petite—Patrie boroughs came together to create a pilot project zone in Montréal where priority is given to self-driving electric vehicles.

The self-driving vehicles, privately owned or carsharing (Communauto, Car2go and the new Goober pods) as well as self-driving STM shuttles travel at a speed of 25 km/h to ensure maximum security for users, cyclists and pedestrians ("Zero accident" policy from the City). This policy ensures fluid circulation without traffic jams, with dynamic traffic lights thanks to a network of connected sensors. All this gives users the freedom to take part in activities such as working, writing, or listening to music in their vehicle without being disturbed by jerking movements. Vehicles with drivers must adapt to these speeds, or risk deterrent fines. The new self-driving traffic regulation centre (SDTRC) does, however, authorize a speed of 50 km/h during morning and evening peak hours on certain major roads, such as Papineau Avenue, Iberville Street and Saint-Joseph Boulevard. To ensure the safety of pedestrians and prevent them from crossing these roads in on a whim, safety barriers have also been erected along these roads.

Samia, 30, lives in Rosemont. She's a massage therapist, strongly suited to therapeutic relationships and an animal rights activist. She lives with her partner, Robin, a computer technician, and her cat, Linus, 4. As often as possible, she lets Linus roam freely throughout town, as she can always track him thanks to his connected collar. The very moderated speed of the self-driving cars reassures her about her cat. Furthermore, she appreciates that in this Montréal pilot project zone, the cars are set in "altruistic" mode, which means they act in the interests of the greatest number of people, even at the expense of the person in the car.

But during the summer, a group of cyclists grows tired of seeing the many safety barriers that appropriate public space for self-driving cars. Since the end of August, they have been protesting by organizing "free bike parades" on the borough's streets for the sake of sharing the road with all eco-friendly methods of transportation, never hesitating to throw themselves under the wheels of the self-driving vehicles, knowing that their "altruistic setting" saves them from danger. But on this October morning, Samia, in her car, doesn't know that her husband Robin modified—out of love—the setting in her car to make it "selfish": it now preserves the driver's interests in the event of an accident. When Laurène, a free bike activist, jumps the security barrier and throws herself in front of the car on Papineau Boulevard, it does not react as planned. An accident occurs that severely injures Laurène, because the CRTA technicians didn't lower the speed from 50 km/h to 10 km/h when she jumped the safety barrier. Samia is in a state of shock.

Theme 3:
PREDICTIVE EDUCATION

Initial scenario:
SELF-DRIVING CAR – SETTING THE ALGORITHM AND SHARING THE ROAD

AUGUST 28, 2025. Carmen starts her third year as a teacher at the Thérèse-Casgrain Elementary School. Like last year, she will be teaching Grade 6. She is eager to use the new teaching methods that the Commission scolaire de la Baie (Baie School Board) has set up a pilot project in the school to improve support for exceptional students and adapt teaching techniques to different learning styles and needs. Last year, Carmen spotted Samuel's learning disabilities a bit late in the school term. Samuel had attention-deficit problems, chatted with his peers instead of listening and would sometimes act aggressively towards friends. Carmen thought his low grades were related to an attention deficit disorder (ADD). She talked about it with Samuel's parents. The conversation did not go very well.

This year everything was going to change thanks to AlterEgo, an artificial intelligence that assists teachers. AlterEgo measures in real time the attention span of students, identifies what hinders their understanding during the lesson and detects exceptional students. The device is very simple: thanks to sensors housed in an electronic bracelet that is connected to the tablet on which the student is working, AlterEgo detects the stress felt by children and when they start to lose focus. The device is also able to analyze reading speeds to identify students with comprehension problems.

Today, Carmen gives the students their bracelet and answers questions from parents who have been invited to attend the first class. The parents were initially a little surprised by the device, but they now seem seduced by everything that it can do. The children play with their electronic bracelet and keep asking AlterEgo questions on their tablet: "AlterEgo, who's your favourite singer?" At the same time, AlterEgo gets acquainted with the students and starts recording the first data.

Carmen explains that her assistant also makes pedagogical recommendations. It can remove parts of the lesson that are deemed ineffective or unsuitable for learning. At the end of the day, Carmen must study AlterEgo's recommendations and each student's profile to plan and adapt the lesson. This greatly improves student tracking. "Thanks to AlterEgo, there's almost no more stress related to exams or evaluating students' needs and progress!" says Carmen. Student assessment will now be almost continuous. However, Carmen is quick to reassure some doubtful parents: teachers will still be assessing students' needs and progress. AlterEgo is an addition to that process. "Who will grade the exams? Will AlterEgo do that too?" asks Hourya's father. Carmen smiles and wraps up her presentation with a joke: "When I have to work at night, I'll definitely need AlterEgo to take care of my kids, Lola and Emiliano. It just might come true!"

Theme 4:
JUSTICE AND PREDICTIVE POLICE

Variable scenario:
PAROLE DECISION

FALL 2025. Sylvia, 29, has been dating Jean for ten years. When she learned Jean cheated on her, she sought revenge by hacking his connected refrigerator.

Knowing Jean's severe peanut allergy, his refrigerator, which would send his grocery list to a partner store, would compile the list according to this information. However, once Sylvia hacked the system, Jean's peanut allergy no longer appeared in the default parameters and the refrigerator produced a list that was no longer adapted to his health requirements. While eating a prepared dish which contained trace amounts of peanuts, Jean started having difficulty breathing and was rushed to the hospital.

Sylvia was arrested for her crime. At the time of sentencing, an algorithm calculated an 80% chance of her relapsing in the next two years, and sentenced her to a two-year prison sentence and a $10,000 fine.

To arrive at this recommendation, the algorithm calculated the risk based on many factors:

> Static historical factors, such as the age at which Sylvia committed her first infraction and her prior offences (Sylvia had already hacked her mother's pillbox at 18, and her neighbourhood's video surveillance camera network at 25);

> Dynamic risk factors: Sylvia's occupation, the company she keeps, her family and romantic relationships, the regret expressed by Sylvia, etc.

Then the algorithm compared Sylvia's case to many other similar cases.

Following the decision rendered by the algorithm, the judge could choose it or order probation for Sylvia, on the condition she follows an all-new rehabilitation program for delinquents that has no data history, meaning no possible interpretation by the algorithm.

The judge, who is keen on social innovation, chose the second option. The rehabilitation program recommends Sylvia be evaluated and follow a routine personalized plan for two and a half years, as well as find a legal job. Given her hacking skills, Sylvia is also asked to put her knowledge to good use by contributing to the field of cyber security.

Theme 5:
WORKPLACE

Initial scenario:
SOCIALLY RESPONSIBLE RESTRUCTURING

JANUARY 15, 2025. Created in 2020 in Montréal, Zéro Carbone Logistique (ZCL) is a new world leader in sustainable logistics, and has witnessed incredible growth over the past five years. The company currently employs 3000 people in Montréal.

Ever since it was founded, ZCL has wanted to include its environmental and social objectives in its shareholder agreement by adhering to B Corp status[31] and by following the ISO 26000 standards on the social responsibility of companies. This policy was beneficial for ZCL because many union funds and socially responsible investment funds quickly invested in the company, which became a poster child for green start-ups in Quebec.

However, ZCL is a company that must be profitable, and it faces very fierce competition when it comes to the cost of services: offering environmental value isn't enough to prosper. Like many companies, it conducted a financial audit and the report strongly recommended a radical scenario to ensure the company's sustainability: massively investing in AI and the automation of several tasks, starting in 2020. This includes calculating each trip's carbon footprint, self-driving electric trucks, parcel sorting, routing blimps and electrical boats, as well as administrative follow-up on files. In total, 1000 jobs out of 3000 could be eliminated, and 1000 others must move towards types of cooperation between humans and co-bots! For ZCL management, there's no way this evolution will be done abruptly, and they wish to establish a "socially responsible restructuring", by carefully preparing the collaborators for new positions.

[31] A certification issued to companies that satisfy societal, environmental, governance and public transparency requirements.

Nabila, one of the founders of ZCL, suggests the following solution: creating, in partnership with one of the web giants, a massive data processing platform used by AI applications in logistics. Jean-Raymond, the company's union representative, is very worried: he mentions that these companies feed off underpaid workers who spend 15 hours a day coding data to train algorithms, and that it is not a good solution for his colleagues. He would rather establish a cooperative data processing platform. "They have some in California and they're much more in line with our values." But a big web stakeholder is ready to immediately invest in massive data for sustainable logistics and create, with ZCL, a subsidiary in Montréal that could hire most of the 1,000 people. Time is running out; their investors are pushing for the immediate partnership which is a sure thing, even though it will most certainly have an impact on ZCL's image. Nabila and Jean-Raymond had been raising these issues with the executive committee on many occasions since 2023. They would have liked to seek advice from a public service earlier, but didn't know whom to reach out to and now, it's too late.

< >
Montréal Declaration
Responsible AI_
</ >

PART 2

2018 OVERVIEW OF INTERNATIONAL RECOMMENDATIONS FOR AI ETHICS

TABLE OF CONTENTS

1.	**INTRODUCTION**	**80**
1.1	Methodology	80
1.2	Opening remarks	83
2.	**THEMATIC SUMMARY OF RECOMMENDATIONS**	**84**
3.	**REPORTS ON AI DEVELOPMENT: TECHNICAL DATA SHEETS**	**92**
3.1	The seven reports studied	92
3.2	Reports examined, but not selected	94
3.3	Other reports consulted	97

TABLES AND FIGURES

Table 1: Occurrence of key concepts in the seven documents examined 81

WRITTEN BY

MARTIN GIBERT, Ethics Counsellor at IVADO and researcher in Centre de recherche en éthique (CRÉ)

CHRISTOPHE MONDIN, Research professional for CIRANO

GUILLAUME CHICOISNE, Scientific Programs Director, IVADO

1. INTRODUCTION

In December 2016, Corinne Cath and her colleagues from Oxford University and the Alan Turing Institute published a comparative analysis of artificial intelligence policies from the European Parliament, the House of Commons of the United Kingdom, and the White House[1]. They concluded that these three reports correctly identified various ethical, social and economic issues, but lacked a long-term strategy to develop "good AI". Where do things stand today? How do various government and non-government organizations foresee the changes that AI will bring to society?

Keep in mind that many events have occurred since December 2016 which have changed public and government expectations of AI, and information technologies in general. The first self-driving car crashes have occurred. Revelations on the attempted tampering with the latest American presidential elections via Facebook as well as the Cambridge Analytica scandal which blew up in March 2018, elicited strong reactions and sparked fear for the good health of democracies. Likewise, Google's image has been somewhat tarnished from its veiled collaborations with the American army. We will have a more accurate understanding of the reports analyzed in this document if we put them back into context—this is especially true for the declaration of ethical principles published by Google in June 2018.

1.1

METHODOLOGY

To provide a brief overview of the situation in 2018, we have analyzed seven recently published reports and declarations of principles. The technical sheets on the selected documents are detailed in the third section of this document. We have added files from reports that were examined, but not selected. What initially guided our choice were ethical recommendations, but that is far from always being the case. In fact, much prospective thinking on the future of AI is from a chiefly economic perspective: how, for example, can we develop an ecosystem that fosters innovative AI companies, what is the strategic plan for AI development in a given country? We set aside reports, therefore, that were primarily economic as well as economic recommendations in the reports selected. Moreover, we did not select reports that focussed exclusively on one specific field, such as robotics research ethics or self-driving car regulations. The goal was to examine a general set of recommendations that could be compared to one another.

We also sought a certain diversity when making our selection, to give us a broad enough scope for comparison. That is why two of the reports (Villani and CNIL) are in French, and the other five are in English. One report is from a private company (Google), three are from non-governmental organizations (IEEE, Asilomar and AI Now) and three others present the official policies of several countries (UKRS, Villani and CNIL). Some reports, therefore, are more global in vision, whereas others are more local. Moreover, some reports were relatively concise (Asilomar, Google, AI Now), while others were much longer and detailed, particularly because they included economic considerations.

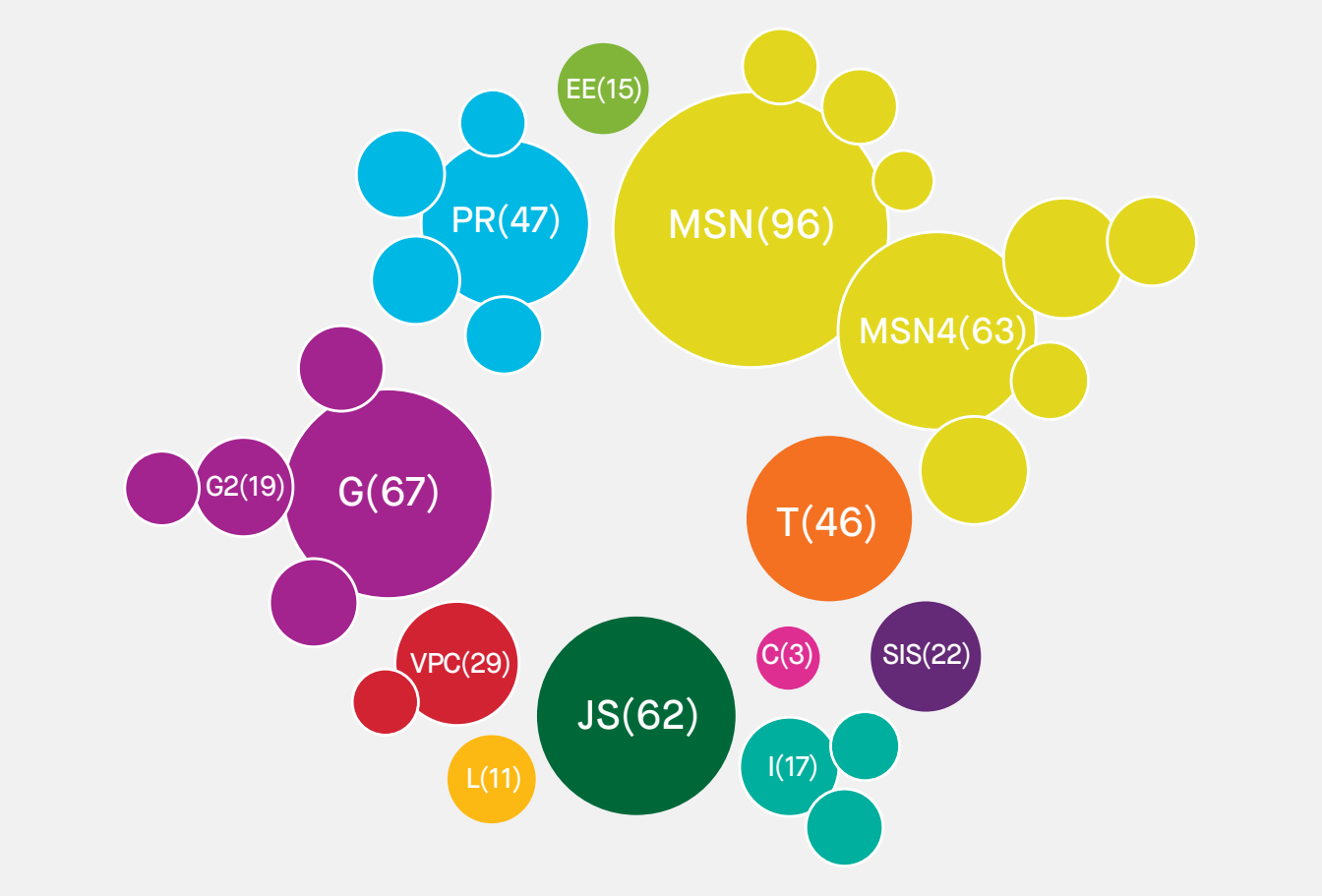
In the technical sheets in section 3, we also highlighted clearly identifiable principles and recommendations. We call "principles" the very general proposals, such as "AI should be beneficial for society", whereas the "recommendations" are more targeted and relatively concrete, such as "we must develop standards to track the source and use of data sets throughout their entire life cycle".

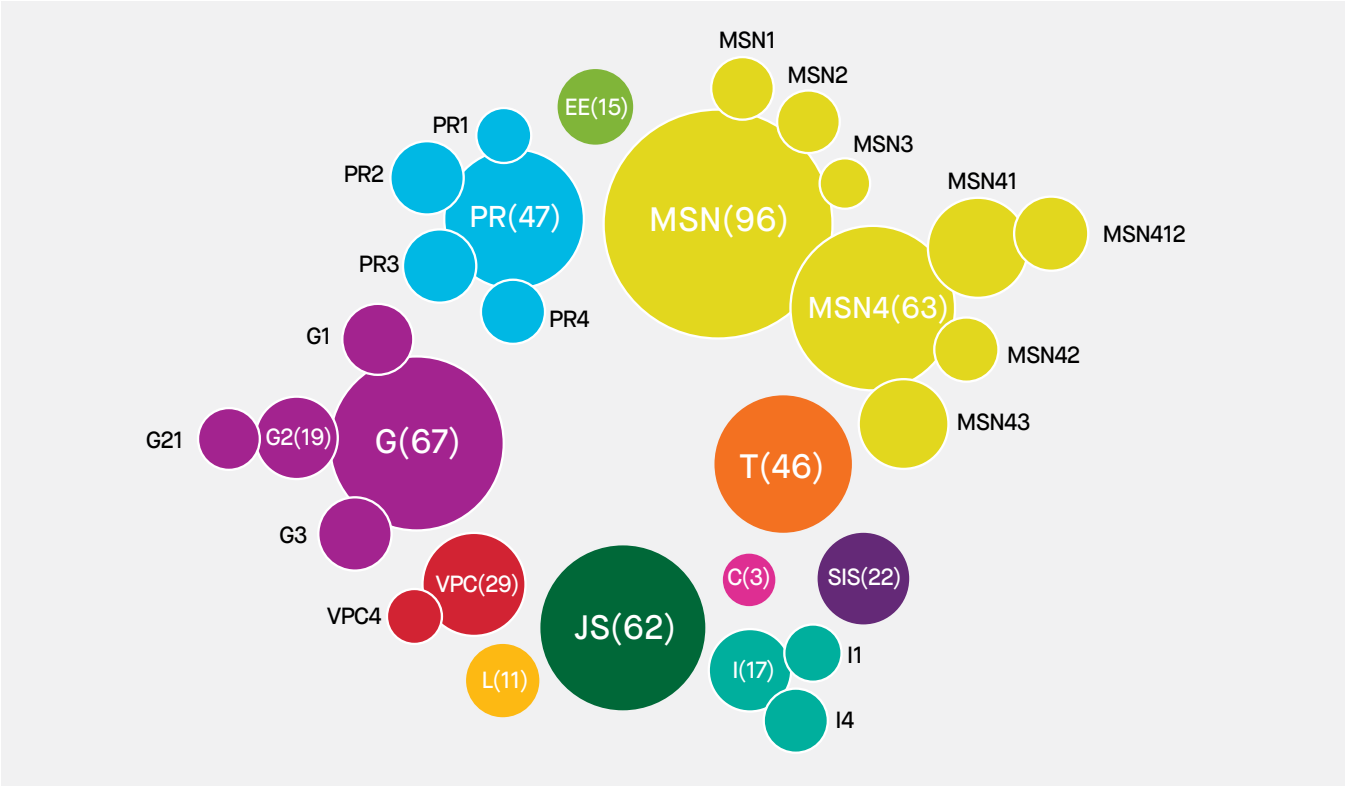
[1] Cath, C., Wachter, S., Mittelstadt, B. et al. Sci Eng Ethics (2018) 24: 505. https://doi.org/10.1007/s11948-017-9901-7

From a methodological standpoint, we started by identifying the ethical recommendations in the seven reports. We retained 230 recommendations. We then classified these recommendations into one of seven categories taken from the preliminary version of the Montréal Declaration for a Responsible Development of Artificial Intelligence: well-being, autonomy, justice, privacy, knowledge, democracy and responsibility—one recommendation may apply to many categories. The advantage of these labels is that they directly refer to what interests us, namely moral values. Of course, classifying a recommendation is often a matter of interpretation and other analysts may have reached different conclusions. We then summed up each value, and presented the results in the second section.

In order to shed new light on the recommendations, we also categorized them according to a set of well-defined key concepts. These concepts are taken from an index developed from citizen recommendations established during the collective reflection (coconstruction) sessions on the Montréal Declaration. This is how we obtained the graphic below:

Table 1 : Occurrence of key concepts in the seven documents examined





LEGEND

C	Consent
EE	Environment and ecology
G	Governance
G1	Collectivism/Individualism
G2	Democratic governance
G21	Digital commons
G22	Citizen participation
G3	Public/private governance
G31	Conflicts of interest
G32	Public institutions/Private companies
G33	Monopoly
I	Influences
I1	Lobbyism
I2	Manipulation
I3	Paternalism
I4	Vulnerability of people
JS	Social justice
L	Freedoms
MSN	Socio-digital mutations
MSN1	Acceptability
MSN2	Activity transformations
MSN3	Respecting humans
MSN4	AI skills
MSN41	Human skills

LEGEND

MSN411	Dependence on technology
MSN412	Digital literacy
MSN413	Transformation of human skills
MSN42	Human-AI synergy
MSN43	AI skills
PR	Sharing responsibility
PR1	Disempowerment
PR2	Accountability
PR3	Shared responsibility
PR4	Decision sovereignty
SAA	Stress, alarmism and anxiety
SIS	Safety and system integrity
T	Transparency
VPC	Privacy and confidentiality
VPC1	Anonymity
VPC2	Confidentiality
VPC3	Right to be forgotten
VPC4	Data ownership
VPC5	Intrusion

1.2

OPENING REMARKS

Before presenting the reports and different summaries per value, we felt a few general remarks were in order. First, the similarities among the reports can be striking: it is often difficult to detect any major divergence among the recommendations from the seven reports. This can partly be explained by research consensus: these reports seek to bring people together, not stir controversy, and they avoid potentially divisive subjects by remaining quite general overall. But it could also be that this convergence simply reflects a fundamental agreement on the types of relationships that we should maintain with AI as a whole. After all, it is hardly surprising that everyone agrees to fight discrimination caused by algorithmic automation, or to promote reinforcing consent when managing user data.

These similarities may also be explained by the fairly homogenous character of the societies these reports are from: rich occidental countries that globally share the same democratic and liberal values. We need, therefore, to address the elephant in the room: how do we regulate AI on an international scale? Data, information and algorithms seem especially impervious to territorial boundaries. What authorities in the United Kingdom, France or any other country can accomplish will always remain very limited, then, in the absence of international cooperation. But is it truly feasible? We must also not forget that calls to reduce discrimination and increase equality exist within a global context of growing inequalities. In other words, it is difficult to isolate issues of AI ethics from issues of international justice.

While similarities exist in the reports we examined, they still contain what could be considered different areas of focus. Some reports highlight political and economic issues (Villani and UKRS) while others concentrate on legal or ethical considerations. Moreover, though they are all presented as reports from experts, the report by CNIL is based, in part, on citizen consultations. The declaration of principles by Google is unique in that it is the only private company represented among these reports. Its declaration could create a potential conflict of interest, but it also is the most likely to have a tangible international impact, given the power of the company.

In terms of content, the most striking difference is in the self-regulation of companies and the role of public bodies in AI system governance. It comes as no great surprise, then, that reports issued by the government, such as the Royal British Society, the "UKRS", or the "Villani report" commissioned by the French government offer more potential solutions from public institutions. They also largely favour legislative tools to meet the challenges the arrival of AI systems heralds—this is also the point of view held by the Institute of Electrical and Electronics Engineers (IEEE). On the contrary, the AI Now and Asilomar reports broach the issue from the perspective of companies that can develop safety tools, self-regulation rules and best practices guides. The CNIL report stands out by suggesting two new principles—vigilance and loyalty of AI systems—while the Villani report pays considerable attention to ecological issues.

Lastly, the pragmatic or prosaic language of these reports is worth mentioning. We are far from the lyricism and existential considerations found in the works of Yuval Harari, Nick Bostrom or sci-fi literature. The focus is not placed on the radical shift that AI is creating in human history, but on a cautious and progressive adaptation of technological innovations. Seen this way, it is worth reiterating the conclusion that Corinne Cath and her colleagues arrived at after reading through the 2016 reports: the general and long-term vision of society with "good AI" is still a work in progress.

2. THEMATIC SUMMARY OF RECOMMENDATIONS

The seven reports or documents quoted in the next section are:

- **AI Now:** the 2017 report AI Now Institute.
- **Asilomar:** the principles that emerged during a Future of Life Institute Conference.
- **CNIL:** the report from the Commission nationale (française) de l'informatique et des libertés.
- **Google:** the principles published by Google in June 2018
- **IEEE:** the report from the Institute of Electrical and Electronics Engineers
- **UKRS:** the report from the British Royal Society
- **Villani:** the report "Donner un sens à l'intelligence artificielle" led by French MP Cédric Villani

WELL-BEING

Every report that we examined contained recommendations explicitly associated with well-being. They appear the most often, which is unsurprising, given that this value is key, and even at times synonymous, with the concept of good. The recommendations associated with well-being are particularly associated with the values of AI skills, social justice, safety and system integrity, privacy and confidentiality, human-AI synergy, and collectivism/individualism.

We note that certain trends start to emerge in the reports. AI Now highlights the challenges of discrimination and biases by demanding, for example, that AI systems that impact society as a whole be developed by people that represent society in all its diversity (AI Now, p. 2). (Villani goes a step further by specifying that every level of the AI design chain must be representative of society [Villani, p. 23].) For its part, CNIL focuses on algorithm loyalty towards people so as to not "betray" them by reinforcing discrimination (CNIL, p. 48). The IEEE puts safety first (IEEE, p. 22) for AI systems, which should always be designed to benefit humans.

Asilomar views it from a research perspective: the goal should not be to create neutral intelligence, but beneficial intelligence (Asilomar); that is why funding should be allocated to this end (Asilomar) and include disciplines such as social sciences, ethics, law, public health or ecology (Asilomar). This is also the case for the UKRS, which demands that the government foster research by developing data sharing standards (UKRS, p. 8) and educate machine learning developers on social and ethical issues (UKRS, pp. 9 and 12). The UKRS also stands out by its focus on research and teaching.

Villani pays special attention to the effects of workplace automation as well as many economic considerations. He recommends, for example, that we create a "public lab for labor transformations" and "launching a legislative reform" (Villani, p. 12) of working conditions in the age of automation. These recommendations fall within a larger project that underscores the general interest and issues of common good, particularly health care: we must develop AI to ensure the "early detection of diseases, the 4 Ps of healthcare [predictive, preventive, personalized and participative], the elimination of medical deserts, emission-free urban transport" (Villani, p. 9). The Villani report is also the only to mention how to promote the ecological transition (Villani, p. 14), which has an obvious impact on well-being.

Google, finally, addresses the notion of well-being in its first principle by stating that AI should be socially beneficial. The principles of the company differ from other reports in that they focus on doing no harm over promoting well-being: it is important, then, to conduct tests to "avoid unjust impacts", limit prejudicial or abusive uses, and not develop potentially destructive technology.

But whose well-being is actually being discussed in these reports? The focus, more or less explicitly, is always on the well-being of humans: IEEE maintains,

for example, that human well-being must be made a priority, using the best available and generally accepted indicators of well-being as a reference point (IEEE, p. 25). No report mentions the well-being of animals. Likewise, when ecological issues are raised (Villani, p. 14) it is from an anthropocentric perspective (as opposed to a pathocentric, biocentric or ecocentric perspective). This does not mean that non-human well-being is not worthy of discussion. In fact, the idea of aligning AI with human values, which is found in Asilomar, leaves the door open to extending compassion towards those who are most vulnerable or concern for other species as human values.

Though they address only human well-being, the reports are "universalist" in that they make no distinction between subcategories of the human population—in other words, it is a question of respecting the universality of human rights. For example, no report claims that only an oligarchy, a state or an organization should benefit from AI—quite the opposite, specifies Asilomar. In other words, as Villani insists, the opportunities associated with the arrival of AI must benefit everyone (Villani, p. 23). He also notes that we must anticipate the impact of technological changes, "which often hit the most fragile portions of the population the hardest" (Villani, p. 14).

When the subject of wealth created by AI (a question that political philosophers call distributive justice) is broached, the reports are careful not to speculate on who should benefit from it. They particularly call for reflection on the matter. Villani recommends "initiating dialogue with industrial partners on how value-added is shared" (Villani, p. 13) while the UKRS advocates that society urgently consider the way "the benefits of automated learning can be distributed among society" (UKRS, p. 12). This "time to reflect" on wealth redistribution is echoed in the fairly common plea in all reports to enhance AI research through collaborations with social sciences or ethics (e.g. Asilomar).

For its part, the preliminary version of the Montréal Declaration suggests the following principle: "AI development should ultimately aim for the well-being of all sentient beings." It takes a more inclusive stance by adopting a pathocentrist view. That is perhaps one of the most original elements of the Montréal Declaration: to consider not only the fate of human beings, but of all individuals that could be affected by AI development.

AUTONOMY

Recommendations explicitly tied to the notion of autonomy are present in every report—with the exception of AI Now. These recommendations are closely linked to issues of human skills, human-AI synergy, AI skills, acceptability, vulnerability of people and social justice.

Overall, the idea that AI must respect human autonomy is defended throughout the different reports. Asilomar states, for example, that AI systems must be designed and operated so they are compatible with the ideals of human dignity, respect of rights, freedoms and cultural diversity. The CNIL (CNIL, p. 57) takes it perhaps one step further: not merely respect autonomy but promote it, starting at the design phase. Among philosophers, this distinction between respecting and promoting generally refers to choosing between a deontological logic of respecting standards (autonomy as a right) and a consequential logic of promoting values (autonomy as a good). However, we must avoid over-interpreting the choice of terms. The CNIL even specifies that it is a matter of correcting a situation since it insists on the importance of "overcoming asymmetries", given that there can be no true autonomy in a situation where one stakeholder holds all the power or all the information. For the CNIL, promoting autonomy is also a question of raising awareness among professionals who use AI (CNIL, p. 55).

Respecting or promoting user autonomy is also expressed in the idea that AI must remain a tool, an instrument that serves users or, broadly speaking, human beings. The IEEE notes that AI systems should always be subordinate to human judgment and control (IEEE, p. 23). This idea echoes the Google principle that AI technologies "must be subjected to appropriate human direction and control" (Google). The CNIL report, incidentally, is named "Comment

permettre à l'homme [sic] de garder la main" (How can humans keep the upper hand?).

This quest for autonomy could be the result of a joint effort between businesses that offer AI and those who use it. For Asilomar, human beings must decide if and how to delegate decision-making to AI systems so they can accomplish goals determined by humans. The CNIL (CNIL, p. 57) offers more concrete recommendations by noting that users should be able to "play with" the parameters of a given system, which has the advantage of fostering understanding. For Google, information and consent must guide how companies use AI, especially "by providing appropriate transparency and control over the use of data", a reminder that the issues of autonomy and privacy are never far away.

Another option appears to be to move away from the tool paradigm by cultivating non-alienating human-machine synergy. For Villani (Villani, p. 18) this synergy could be based on developing innately human skills such as creativity, manual dexterity or problem solving. New ways to reach these types of goals are required: we need new means (Villani, p. 23) or digital literacy training from elementary school through university, for all citizens (CNIL, p. 54).

The CNIL (CNIL, p. 48) proposes a principle of loyalty which sums up the spirit of what sound autonomy management could look like in the age of AI . "A loyal algorithm should not incite, reproduce or reinforce any kind of discrimination whatsoever, even unknowingly, by designers". This loyalty must be understood as not only extending towards individual users, but society as a whole—because all of society could be affected by algorithmic "rulings" that are explicitly unwanted. We also see how issues of autonomy are often aligned with those of justice.

For its part, the preliminary version of the Montréal Declaration suggests the following principle: "AI development should foster the autonomy of all human beings and control the autonomy of computer systems." Because of its very general nature, this principle is in keeping with the other reports. It sets itself apart slightly by introducing the autonomy of computer systems—whereas other reports focus on human autonomy and the risks of it dwindling.

JUSTICE

Every report contains recommendations on justice, with the main themes being social justice, human skills, human-AI synergy, AI skills and respecting humans.

The key idea is that artificial intelligence, and the systems that use its power, must lead to a fairer, more equitable society (AI Now, p. 2). This idea is rooted in two principles:

1. The goal of AI must be to redress the shortcomings of society in these fields (UKRS, p. 12);
2. we must be careful, especially during the development and deployment phases, not to create or perpetuate injustice (Google). These two goals can be reached by providing solutions at many different levels.

AI innovations must benefit everyone (Google). The idea is a trickle-down effect (Villani, p. 16): benefits (in service) and wealth (in knowledge, in technology/technique, in accumulated data) must not be reserved for large private companies (Villani, p. 12) or the upper echelons of society—who may represent a majority of the population in terms of culture, religion or race, or a minority of the population in terms of income, such as the "1%". (Villani, p. 22).

AI innovations must aim for a better world where existing inequalities are addressed and fought in the legal system (Asilomar), in access to health care, or in protecting usually overlooked populations (AI Now, pp. 1 and 2; Villani, p. 18; Google). A national database that helps to objectively identify inequalities between men and women in the workplace (Villani, p. 23) needs to be created to resolve gender-based discrimination issues. Likewise, we must steer AI development toward applications that help improve both economic performance and the common good.

For everyone to benefit from AI, it must be inclusive, at every level (Villani, p. 19). This means that at every stage, from design to deployment to maintenance, an AI system should be examined by public authorities. Incentive policies are also needed to include underrepresented populations such as women or minorities.

Additional training in social sciences and ethics can help sensitize designers to these issues and provide the conceptual and intellectual tools to address them (AI Now, pp. 1 and 2). Likewise, research on algorithm interpretability and robustness as well as issues of equality, privacy and causality must be promoted and funded (UKRS, p. 13).

Lastly, justice also concerns legal institutions that can be directly affected by AI development. Here is what the different reports propose:

> A legal framework must be developed to guarantee social justice, ensure everyone is represented when designing and using algorithms, reduce inequalities, and prevent abuse or misuse that could arise with unregulated AI use (Asilomar).

> An important overhaul of the judicial system on all matters pertaining to artificial intelligence and data is overdue, especially for questions of sovereignty, ownership, data citizenship and governance (UKRS, p. 12; Asilomar; IEEE, p. 22). Likewise, we must give considerable thought to the notion of transparency and its evaluation criteria if we wish to assess the compliance of companies using AI systems (IEEE, p. 30).

> These legal and ethical frameworks should be designed with the buy-in of all stakeholders in society: the scientific community, public authorities, industry players, entrepreneurs and civil society organizations (Villani, p. 21). Control systems should be regularly evaluated to ensure they are satisfactorily fulfilling their mission.

> In the same way that it was determined that a company is a separate legal entity, we must reflect on the legal nature of AI itself (Asilomar).

> When artificial intelligence is involved in legal decisions, auditing, interpretation, verification and explanation measures must be implemented (Asilomar).

For its part, the preliminary version of the Montréal Declaration suggests the following principle: "AI development must promote justice and seek to eliminate discrimination, namely that of gender, age, mental and physical abilities, sexual orientation, ethnic and social origins and religious beliefs."

This statement primarily addresses social justice and problems of equality and equity whether by redressing past discrimination or anticipating future discrimination. The Montréal Déclaration does not specify how to achieve these goals, unlike many reports that suggest, for example, more inclusion and social representation in the early phases of designing artificial intelligence systems. Moreover, it does not discuss implications that are specific to the legal system.

PRIVACY

Explicit recommendations on privacy are contained in every report, with the exception of AI Now. These recommendations are namely associated with issues of privacy and confidentiality, collectivism/individualism, digital commons, governance, social justice, transparency, safety and system integrity.

On a very general level, the issue of privacy expresses the idea that the user should have control over their data—a link can therefore be made with autonomy. Asilomar, for example, maintains that people should have the right to access, manage and control the data they generate, while Google claims that protection of privacy should play an important role in the design of AI principles and AI system development. We note, however, that the reports provide few details as to general privacy principles. The issue appears difficult to address in such a general manner.

Protection of privacy implies various governance frameworks, namely regulatory and standard-setting bodies (IEEE, p. 22). For the CNIL (CNIL, p. 45), the law is responsible for overseeing the use of personal data by AI. A pertinent example is offered by (Villani, p. 11) who, in the wake of the General Data Protection Regulation (GDPR), mentions the right to data portability, meaning individuals' rights to recover the data they generate on one platform and use on another platform.

Two trends seem to emerge in the socio-political models that determine data governance. Villani and the CNIL seem to adopt the logic of data as a

common good, while the UKRS appears to align itself with a more "liberal" logic, or at least one more centred on the individual. Once again, we need to be careful in contrasting these approaches as it is difficult to infer a general trend from a few recommendations. Villani (Villani, p. 11) asks government authorities impose "openness on certain data of public interest". We need only think of medical data that, when pooled together, could help advance research and benefit an entire population, or environmental data, for example, which could help collectively fight climate change. This suggestion echoes that of the CNIL (CNIL, p. 59), which suggests that the state launch a "major research program based on data contributed by citizens exercising their right to data portability among private stakeholders."

For the UKRS, protecting privacy in scientific research takes precedence. The issue is of protecting individuals, which is why researchers should keep track of potential future uses for the data they collect, and integrate this aspect into the consent participants provide for research (UKRS, p. 8). This concern must be present from the moment data is collected until it is potentially shared or redistributed. The contrast between the two types of logic is not that significant, as the CNIL also suggests developing research infrastructure that is "respectful of personal data" (CNIL, p. 59), while the UKRS is not opposed to the logic of a "data commons" when it is generated by studies funded by public funds or charities (UKRS, p. 8).

We conclude by noting that there is an evident link between the issues of protection of privacy and justice since personal data could serve as the basis for discriminatory policies. This aspect is present in most reports.

For its part, the preliminary version of the Montréal Declaration proposes the following principle: "AI development should guarantee the respect of privacy and allow those who use it to access their personal data as well as the kinds of information used by the algorithm." Although this principle provides a summary that reflects the findings of other reports, it does not provide an exhaustive overview of the complex and wide-ranging subject of privacy. In particular, this principle of the Montréal Declaration does not broach issues of transparency which, in many ways, are the corollary of privacy—and are analyzed in the next section on knowledge.

KNOWLEDGE

Recommendations on knowledge are in every report. The most common themes are social justice, transparency, human skills and digital literacy.

The two main areas of focus are increasing knowledge of the public and the authorities that will validate or verify AI systems. The autonomy, and transparency, of public and governing bodies can only exist if we let the public and the government exercise it by providing the necessary mechanisms and infrastructure as well as training, education and critical thinking.

A new digital literacy must be developed to hone critical thinking and understanding of these new technologies (CNIL, p. 54), from grade school through university, for all citizens. To achieve this, we must encourage new ways of thinking about intellectual autonomy and get people involved in thinking about everyday AI issues (CNIL, p. 57)—for example, understanding what it means when you give your consent. In other words, we must address the asymmetry that exists between AI service providers and users/citizens.

To protect the public, we need to think about possible ways AI systems could be used for harm. This implies establishing the framework for an educational method and appropriate measuring tools (IEEE, p. 31); for example, a validation test in schools. Again, notions of ethics and social sciences are suggested to complement this learning (IEEE, p. 31). People who are most "at risk", meaning those identified as being especially gullible and/or who would suffer the worst consequences from harmful uses are those to be targeted as priorities (IEEE, p. 31).

A number of recommendations, in addition to those intended for the public, are addressed to government agents, elected representatives who vote on laws,

the legal system that will enforce them and the institutions that will serve as safeguards (IEEE, p. 31). Other "at risk" sectors, such as medicine, human resources (recruitment) or even marketing should be especially vigilant (CNIL, p. 55).

Obviously, algorithm and AI system designers are also affected by these measures: their training should be complemented by studies in humanities so that they grasp the social and economic issues of the solutions they come up with and understand the impact their solutions may have in practice (CNIL, p. 55). Many reports recommend reinforcing cultural, social and gender diversity; the idea is that by multiplying representatives of society at every level of AI design, we will have a better understanding of all the parameters, context and viewpoints to take into consideration (CNIL, p. 55).

Numerous recommendations underscore the knowledge required to correctly operate the AI control systems and evaluation infrastructure. We first need to establish standards and regulatory bodies to monitor the various steps of the design process for AI systems and ensure they respect human rights, freedoms, dignity, privacy and traceability (IEEE, p. 22). These standards must be implemented by public institutions (IEEE, p. 30) that develop transparent measuring tools that are accessible to the public (AI Now, p. 1) and designed by impartial experts and professionals.

One recurrent recommendation in these reports is the need for transparent regulatory bodies. Giving the public access to all these evaluation methods will allow them not only exercise, but demonstrate their knowledge. With trained and motivated users and systems led by transparent and authoritative committees, the last step would be to leave citizens free to experiment, deploy their digital literacy and exercise their critical thinking. It suggests, for example, that the various user platforms for AI systems offer information on how their algorithms operate (CNIL, pp. 45 and 48). Specific information on the data used and algorithm logic could be made available on user profile pages (CNIL, p. 56). To promote understanding, users should be able to "play" with the system by changing parameters (CNIL, p. 57).

One last point: we must ensure the transition by verifying and improving training in schools. Digital literacy is defined in different ways, from the aforementioned ethics and critical thinking to the knowledge of key principles of programming or machine learning (UKRS, p. 9). Once again, it is a matter of addressing the information asymmetry that can exist between users, developers and citizens. Governments, experts in mathematics and programming, companies and education professionals must all contribute to this digital literacy in order to build a much-needed and sufficient knowledge base (UKRS, p. 9). Many recommendations highlight the importance and interest of addressing notions of ethics, social sciences and public health in educational activities (UKRS, p. 9).

As for the educational system, its mission should also be to train a new generation of workers and researchers with the skills required to navigate a world of AI systems. Not only should we reconsider the initial training offered in university, but we must also provide ongoing training and new skill sets to workers whose tasks will be drastically altered. These recommendations are all the more meaningful in the context of job insecurity caused by machines replacing humans (UKRS, p. 9). Both universities and industries must reflect on future needs in terms of skills, from machine learning to the science of data (UKRS, p. 9).

On the subject of knowledge, the preliminary version of the Montréal Declaration proposes the following principle: "AI development should promote critical thinking and protect us from propaganda and manipulation." If awakening critical thought echoes the notions of digital literacy found to varying degrees and ways in the reports, the Montréal Declaration focuses on protecting the public from propaganda and manipulation, whereas the notions of accomplishment, freedom, and power are more evident in other reports. For many, knowledge not only offers protection, but opens the door to many possibilities.

DEMOCRACY

The value (or notion) of democracy is apparent in all reports, in comparable proportions. The recommendations that address democracy are associated with governance, collectivism/individualism, democratic governance, digital commons, privacy and confidentiality, in particular.

The first theme deals with governance. As we have already seen with the principle of autonomy, the reports insist that AI remain under human control (AI Now, p. 1), which explains the need for a specialized supervision framework (IEEE, p. 22; UKRS, p. 12) or audit systems (CNIL, p. 57). Can we allow the private sector to self-regulate? The answer that emerges from these reports is rather pessimistic—but we must admit that the opposing viewpoint is essentially nonexistent, as the only company whose principles/recommendations are available (Google) does not address the question. As for the type of governance, certain recommendations use a relatively classic top-down logic: IEEE or UKRS, for example, with the idea of seeking social acceptability or "consulting" citizens (IEEE, p. 31). Asilomar discusses the need for dialogue between researchers and policy-makers. The more radical or direct conceptions of democracy do not appear explicitly in the reports.

Regardless, everyone agrees that AI development must be regulated—Villani even specifies that, for example, a special framework must be developed to protect the most sensitive sets of data (Villani, p. 20). But what lends these recommendations a truly democratic dimension is that the framework or control in question must be transparent. AI Now advocates that AI systems used by public agencies be subject to public audits, tests and revisions (AI Now, p. 1). The idea of a "public body of experts" that would control the algorithms "to verify for example that "they do not discriminate" is also echoed by the CNIL (CNIL, p. 58), which goes even further than AI Now because the mission of these experts does not appear limited to the public sector. The issue of algorithmic opacity is often mentioned, since it hinders transparency. Villani notes that being able to "open the black boxes" is a democratic issue (Villani, p. 21). Support for research in the field of algorithm explainability, therefore, is necessary (Villani, p. 21).

Democracy is also present in calls for diversity—cultural, social and gender, as specified by CNIL (CNIL, p. 55)—among algorithm developers since it is unlikely that a sub-group (usually of rich white men) can adequately anticipate and respond to the needs of all members of society. Villani, therefore, promotes "inclusive and diverse" AI (Villani, p. 22) whereas the IEEE (IEEE, p. 27) recommends that designers and developers be aware of the diverse cultural norms that exist among AI system users. Finally, for Google, one of the roles of companies is to share knowledge and thereby democratize AI so that more people can develop useful applications for it (Google).

The preliminary version of the Montréal Declaration proposes the following principle: "AI development should foster informed participation in public life, cooperation and democratic debate." The lack of issues on diversity is unsurprising because they are addressed in the principle of justice in the Montréal Declaration. It is worth considering, however, whether issues of governance and transparency should accompany this principle. The principle of democracy in the Montréal Declaration says nothing about who should control AI development and how sharing between public and private governance, experts and laypeople should be addressed.

RESPONSIBILITY

All reports contained recommendations on responsibility. The most predominant themes were safety and system integrity, social justice, AI skills, sharing responsibility, accountability and shared responsibility.

The issue of decision-making first touches upon the notion of responsibility: when AI can act alone, when should it be supervised or completed by a human being (AI Now, p. 1)? For some, a machine must never make a decision by itself (meaning with no human intervention) if there are serious consequences for people (CNIL, p. 45).

To correctly assign responsibility to one entity or the other (or both), we must ensure that the humans who interact with AI have the necessary

training to understand, think critically and measure the limits and biases that need to be corrected. Some recommendations go one step further by suggesting that the moment AI shows biases and discrimination, and as its presence in our social and economic lives increases, "opening the black box" becomes a democratic issue (Villani, p. 18). Because competitiveness issues foresee companies not all nor always being completely transparent, it is often recommended that the use of AI in the public sphere be as transparent as possible. First, by not relying on private companies to manage public systems (AI Now, p. 1), then by subjecting public systems to the strictest tests, evaluations, audits, inspections, and responsibility standards (AI Now, p. 1).

Being responsible also means anticipating problems: how can we avoid hurdles, what infrastructures should we implement? Some recommendations on this issue are clear: the ruling principle should be vigilance (CNIL, p. 50) and AI designers should always remember that algorithms can be unpredictable and they are autonomous and constantly evolving. This principle of vigilance seeks to temper, or at least counterbalance the risk of excessive trust in AI (CNIL, p. 50). Many potential solutions are proposed, such as devising recording and traceability systems to go back to the source of an algorithm and determine responsibility in case of a problem (IEEE, p. 27).

Each report highlights the fact that currently, the legal system can barely keep up with the frenzied pace of data and AI development, and consequently, with ways to regulate these new technologies. Resources need to be mobilized so that it can stay apace (Asilomar).

Two key elements appear necessary to guide AI systems:

1. Involving the judicial system to control, correct, delineate and help;
2. involving independent scientists in the design of devices to monitor, call and label all these AI systems. Both groups will have to work together to establish good control test practices (Asilomar).

Companies must not remain passive, however. Since they must ensure not to amplify biases or make mistakes (AI Now, p. 1), a significant share of the work to be done is in prevention, namely by using test versions before the global launch of an AI app (AI Now, p. 1). These preliminary tests must not only verify the way the algorithms are built but, above all, verify the data on which they were trained (AI Now, p. 1). This is why it is recommended to have information on the source and management of this training data, as well as back-ups so they can be explored in case of an anomaly (AI Now, p. 1).

Responsibility in the legal field is a hot topic, and the responsibility for making the most appropriate decision and preventing injustice (creating it, reinforcing it) is at the heart of many discussions. Hence, Asilomar recommends that any autonomous system involved in legal decisions be able to provide clear explanations on its decision-making process (Asilomar). These explanations should be analyzed by a competent person with adequate training to understand the workings of the algorithm and offer an intelligible explanation.

The theme of responsibility also concerns mediation between the public and AI system providers, therefore openness and transparency. We must include all members of society in the debate on human responsibility (Villani, p. 22). The public will be asked to think critically in mediation cases (previously discussed)—if the public wishes to defend itself in case of a dispute, the algorithms must be explainable, and the public must be able to understand them—, as well as in public and citizen consultations or open national audits.

For its part, the preliminary version of the Montréal Declaration proposes the following principle: "The different stakeholders in AI development should assume responsibility by minimizing the risks of these technological innovations." The Montréal Declaration sums up the essence of the recommendations made in the various reports, but remains very general (as the reports can offer more detailed recommendations). The Montréal Declaration, therefore, could help clarify the role of the many different stakeholders involved in building these systems, the impact of their work on each other and the pitfalls that should be avoided.

3. REPORTS ON AI DEVELOPMENT: TECHNICAL DATA SHEETS

3.1 THE SEVEN REPORTS STUDIED

(AI NOW) AI NOW 2017 REPORT

Subtitle: no
Published: November 2017
Country: USA
Language: English
Organization or signatories: AI Now Institute (report signed by Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, Kate Crawford)
Number of pages: 37
Summary: yes (3 pages)
Well identified general ethical principles: no
Well identified recommendations: yes (10)
Main themes: work and automation, biases and inclusion, rights and freedoms, ethics and governance.
Notes: An annual report that quotes many recent studies and is devoted to updating people on advances in research.
Link: https://ainowinstitute.org/AI_Now_2017_Report.pdf

(CNIL) HOW CAN HUMANS KEEP THE UPPER HAND? THE ETHICAL MATTERS RAISED BY ALGORITHMS AND ARTIFICIAL INTELLIGENCE

Subtitle: Report on the public debate led by the French data protection authority (CNIL) as part of the ethical discussion assignment set by the Digital Republic bill
Published: December 2017
Country: 80
Language: English (translated from French : Comment permettre à l'homme de garder la main - Les enjeux éthiques des algorithmes et de l'IA)
Organization or signatories: CNIL: Commission nationale informatique et liberté (foreword by Isabelle Falque-Pierrotin, president of the CNIL)
Number of pages: 80
Summary: yes (2 pages)
Well identified general ethical principles: yes (vigilance and loyalty)
Well identified recommendations: yes (6)
Main themes: ethical uses of AI, applications for each field (health care, education, living in society and politics, culture and media, justice, banks and finance, safety and defence, insurance, work and HR).
Notes: One of the most thorough reports on the ethical issues of AI.
Link: https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf

(IEEE) ETHICALLY ALIGNED DESIGN. VERSION 2—FOR PUBLIC DISCUSSION

Subtitle: A vision for prioritizing human well-being with autonomous and intelligent systems.
Published: December 2017
Country: international
Language: English
Organization or signatories: IEEE (Institute of Electrical and Electronics Engineers); signed by IEEE subcommittees that regroup several hundred international participants.
Number of pages: 266
Summary: yes (17 pages)
Well identified general ethical principles: yes (5)
Well identified recommendations: yes
Main themes: ethical, legal, political issues; questions specifically tied to information and communication technologies; safety; ethics by design; data control.
Notes: Each chapter was written by committees of experts.
Link: https://ethicsinaction.ieee.org/

(ASILOMAR) ASILOMAR AI PRINCIPLES

Subtitle: no
Published: 2017
Country: international
Language: English with Chinese; German, Japanese, Korean and Russian translations available
Organization or signatories: Future of Life Institute, signed by over 1200 researchers and 2500 non-researchers.
Number of pages: 2
Summary: no
Well identified general ethical principles: yes (23)
Well identified recommendations: no
Main themes: ethics of research, moral values, long-term issues.
Notes: It is not a report, but a series of principles that stem from discussions between experts during a conference in Asilomar, California. In 1975, another conference in Asilomar established bioethics principles.
Link: https://futureoflife.org/ai-principles/?cn-reloaded=1

(UKRS) AI IN THE UK: READY, WILLING, AND ABLE?

Subtitle: no
Published: April 16, 2018
Country: UK
Language: English
Organization or signatories: Parliament (House of Lords); 13-person committee.
Number of pages: 184
Summary: yes (5 pages)
Well identified general ethical principles: no
Well identified recommendations: yes (73)
Main themes: questions of ethics and economics policy ("innovation in AI"). Impact of AI on different fields: economy, work, education, health care, justice.
Notes: The report is divided into 420 paragraphs, with the author usually identified in the notes.
Link: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

(VILLANI) DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

Subtitle: Pour une stratégie nationale et européenne
Published: March 8, 2018
Country: France
Language: French
Organization or signatories: Parliamentary missions entrusted to MP Cédric Villani and six (6) other members of parliament.
Number of pages: 235
Summary: yes (15 pages)
Well identified general ethical principles: no
Well identified recommendations: no
Main themes: questions of ethics and political economy, research policies, impact on work and education sectors, health, agriculture, transportation and defence.
Link: http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/184000159.pdf

(GOOGLE) AI AT GOOGLE: OUR PRINCIPLES

Subtitle: no
Published: June 7, 2018
Country: USA
Language: English
Organization or signatories: Google, presented by its CEO Sundar Pichai
Number of pages: 3
Summary: no
Well identified general ethical principles: yes (7)
Well identified recommendations: yes (4)
Main themes: AI ethics
Notes: The company commits to not deploying AI in certain fields (weapons) or circumstances (against human rights).
Link: https://www.blog.google/technology/ai/ai-principles/

3.2

REPORTS EXAMINED, BUT NOT SELECTED

A NEXT GENERATION ARTIFICIAL INTELLIGENCE DEVELOPMENT PLAN

Subtitle: no
Published: July 2017
Country: China
Language: English (translation)
Number of pages: 28
Summary: no
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/

STRATEGY FOR DENMARK'S DIGITAL GROWTH

Subtitle: no
Published: 2018
Country: Denmark
Language: English
Organization or signatories: Ministry of Industry, Business and Financial Affairs
Number of pages: 68
Summary: yes (6 pages)
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: https://em.dk/english/news/2018/01-30-new-strategy-to-make-denmark-the-new-digital-frontrunner

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS

Subtitle: Artificial Intelligence for Europe
Published: April 25, 2018
Country: European Union
Language: English
Organization or signatories: European Commission
Number of pages: 20
Summary: no
Well identified general ethical principles: yes
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

FINLAND'S AGE OF ARTIFICIAL INTELLIGENCE

Subtitle: Turning Finland into a leading country in the application of artificial intelligence: Objective and recommendations for measures
Published: December 18, 2017
Country: Finland
Language: English
Organization or signatories: Ministry of Economic Affairs and Employment
Number of pages: 76
Summary: yes (3 pages)
Well identified general ethical principles: no
Well identified recommendations: yes (8)
Main themes: national strategy for economic development
Link: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y

ETHICS COMMISSION AUTOMATED AND CONNECTED DRIVING

Subtitle: no
Published: June 2017
Country: Germany
Language: English
Organization or signatories: Federal Ministry of Transport and Digital Infrastructure
Number of pages: 36
Summary: no
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: Ethics of self-driving vehicles
Link: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

NATIONAL STRATEGY FOR ARTIFICIAL INTELLIGENCE #AIFORALL

Subtitle: Discussion paper
Published: June 2018
Country: India
Language: English
Organization or signatories: NITI Aayog
Number of pages: 115
Summary: yes (3)
Well identified general ethical principles: yes
Well identified recommendations: yes
Main themes: national strategy for economic and societal development
Link: http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

ARTIFICIAL INTELLIGENCE AT THE SERVICE OF CITIZENS

Subtitle: no
Published: March 2018
Country: Italy
Language: English
Organization or signatories: The Agency for Digital Italy
Number of pages: 79
Summary: yes (5 pages)
Well identified general ethical principles: yes
Well identified recommendations: yes
Main themes: AI's impact on society and the public administration to promote change
Link: https://ia.italia.it/en/assets/whitepaper.pdf

ARTIFICIAL INTELLIGENCE TECHNOLOGY STRATEGY

Subtitle: Report of Strategic Council for AI Technology
Published: March 31, 2017
Country: Japan
Language: English
Organization or signatories: Strategic Council for AI Technology
Number of pages: 25
Summary: no
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: http://www.nedo.go.jp/content/100865202.pdf

TOWARDS AN AI STRATEGY IN MEXICO

Subtitle: Harnessing the AI Revolution
Published: June 2018
Country: Mexico
Language: English
Organization or signatories: British Embassy in Mexico through the Prosperity Fund, Oxford Insights, C Minds
Number of pages: 52
Summary: yes (3 pages)
Well identified general ethical principles: no
Well identified recommendations: yes (21)
Main themes: national strategy for economic development
Link: https://docs.wixstatic.com/ugd/7be025_e726c582191c49d2b8b6517a590151f6.pdf

SHAPING A FUTURE NEW ZEALAND

Subtitle: An Analysis of the Potential Impact and Opportunity of Artificial Intelligence on New Zealand's Society and Economy
Published: May 2018
Country: New Zealand
Language: English
Organization or signatories: AI Forum of New Zealand
Number of pages: 108
Summary: yes (5 pages)
Well identified general ethical principles: yes
Well identified recommendations: yes (14)
Main themes: national strategy for economic development
Link: http://resources.aiforum.org.nz/AI+Shaping+A+Future+New+Zealand+Report+2018.pdf

ARTIFICIAL INTELLIGENCE IN SWEDISH BUSINESS AND SOCIETY

Subtitle: Analysis of development and potential
Published: May 2018
Country: Sweden
Language: English
Organization or signatories: Vinnova
Number of pages: 32
Summary: no
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: economic development and public services
Link: https://www.vinnova.se/contentassets/29cd313d690e4be3a8d861ad05a4ee48/vr_18_09.pdf

INDUSTRIAL STRATEGY

Subtitle: AI Sector Deal
Published: April 2018
Country: UK
Language: English
Organization or signatories: Government
Number of pages: 21
Summary: yes (3 pages)
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4_.pdf

PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

Subtitle: no
Published: October 2016
Country: USA
Language: English
Organization or signatories: Executive Office of the President, National Science and Technology Council Committee on Technology
Number of pages: 58
Summary: yes (4)
Well identified general ethical principles: yes
Well identified recommendations: yes (23)
Main themes: current state of AI, present and future applications, questions raised for society
Link: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

Subtitle: no
Published: October 2016
Country: USA
Language: English
Organization or signatories: National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee
Number of pages: 48
Summary: yes (2 pages)
Well identified general ethical principles: no (a few)
Well identified recommendations: yes (7)
Main themes: objectives for AI research funded by the federal government
Link: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf

ARTIFICIAL INTELLIGENCE, AUTOMATION, AND THE ECONOMY

Subtitle: no
Published: December 2016
Country: USA
Language: English
Organization or signatories: Executive Office of the President
Number of pages: 55
Summary: yes (4 pages)
Well identified general ethical principles: no
Well identified recommendations: yes (3)
Main themes: impact of AI automation on the economy and strategies to increase the benefits
Link: https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

SUMMARY OF THE 2018 WHITE HOUSE SUMMIT ON ARTIFICIAL INTELLIGENCE FOR AMERICAN INDUSTRY

Subtitle: no
Published: May 10, 2018
Country: USA
Language: English
Organization or signatories: The White House Office of Science and Technology Policy
Number of pages: 15
Summary: yes (1 page)
Well identified general ethical principles: no
Well identified recommendations: yes
Main themes: national strategy for economic development
Link: https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf

3.3

OTHER REPORTS CONSULTED

(Sweden) National Approach for Artificial Intelligence
https://www.regeringen.se/49a828/contentassets/844d30fb0d594d1b9d96e2f5d57ed14b/2018ai_webb.pdf

(Germany) Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz
https://www.bmwi.de/Redaktion/DE/Downloads/E/eckpunktepapier-ki.pdf?__blob=publicationFile&v=4

(Finland) Work in the Age of Artificial Intelligence
http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160931/19_18_TEM_Tekoalyajan_tyo_WEB.pdf

(China) Three-Year Action Plan to Promote the Development of New-Generation Artificial Intelligence Industry
http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5960820/content.html

(Australia) Australia 2030: Prosperity Through Innovation
https://www.industry.gov.au/sites/g/files/net3906/f/May%202018/document/pdf/australia-2030-prosperity-through-innovation-full-report.pdf

(Thanks to Paloma Fernandez-McAuley for her help.)

< >
Montréal Declaration
Responsible AI_
</ >

PART 3

SUMMARY REPORT OF RECOMMENDATIONS FROM THE WINTER CO-CONSTRUCTION WORKSHOPS

TABLE OF CONTENTS

1.	**SUMMARY**	**100**
2.	**CO-CONSTRUCTION DATA: EXPLANATORY REMARKS**	**102**
3.	**MAIN DIRECTIONS EXPECTED BY CITIZENS**	**104**
4.	**CITIZEN PERCEPTION OF RESPONSIBLE AI DEVELOPMENT ISSUES**	**106**
4.1	Introduction	106
4.2	Main categories of risks and issues in responsible AI development	110
5.	**POTENTIAL SOLUTIONS AND FRAMEWORK FOR RESPONSIBLE AI DEVELOPMENT**	**125**
5.1	Introduction	125
5.2	Education	127
5.3	Legal system and predictive policing	131
5.4	Workplace	135
5.5	Healthcare	140
5.6	Smart cities and connected objects	144
6.	**CONCLUSION**	**149**

TABLE AND FIGURES

Table 1: Potential solutions proposed to respond to the issues identified	101
Table 2: Priorities identified by citizens according to the principles of the Declaration (number of tables)	106
Table 3: Mind map of the issues	111
Table 4: Three key potential solutions at all tables	126
Table 5: Potential solutions or general guidelines for the education sector	127
Table 6: Potential solutions or general guidelines for the legal system and predictive policing sector	131
Table 7: Potential solutions or general guidelines for the workplace sector	135
Table 8: Potential solutions or general guidelines for the healthcare sector	140
Table 9: Potential solutions or general guidelines for the smart city and connected objects sector	144

WRITTEN BY

NATHALIE VOARINO, Scientific Coordinator, PhD Candidate in Bioethics of Université de Montréal

CAMILLE VÉZY, Doctoral student in communication, Université de Montréal

VALENTINE CROSSET, Doctoral student in criminology, Université de Montréal

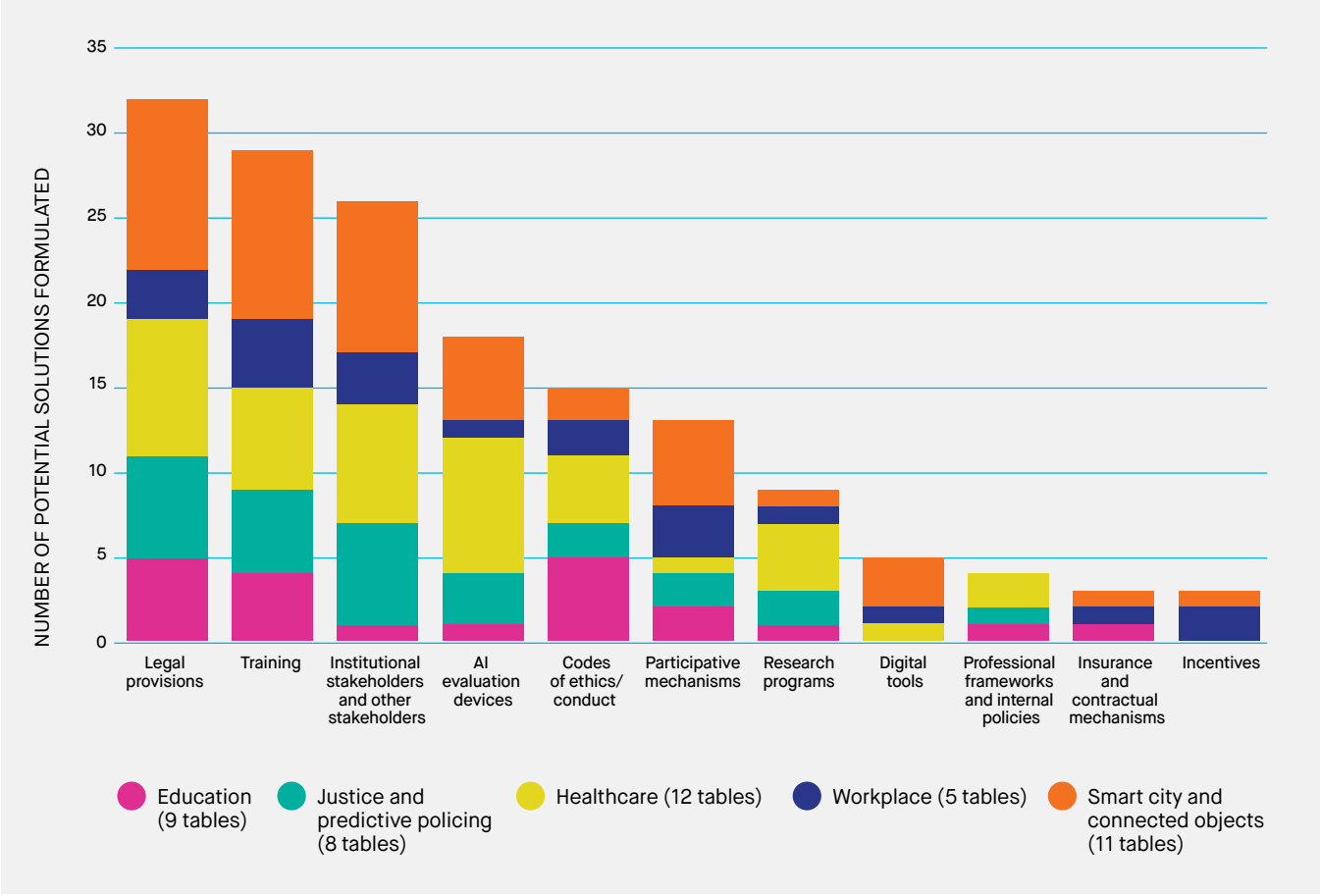
ALESSIA ZARZANI, Ph.D in Planning, Université de Montréal and Ph.D in Landscape and Environment, Université la Sapienza de Roma

1. SUMMARY

Citizens met at 45 tables to discuss their perception of risks and issues in the responsible development of artificial intelligence (AI).

They formulated 11 categories of potential solutions to these risks and issues:

Table 1: Potential solutions proposed to respond to the issues identified



2. CO-CONSTRUCTION DATA: EXPLANATORY REMARKS

The current section explains the results collected during the co-construction tables held in winter 2018 for the Montréal Declaration. There were 45 discussion tables in total that brought together hundreds of citizens. Discussions were held around five major sectors of AI development: education (9 tables); justice and predictive policing (8 tables); healthcare (12 tables), workplace (5 tables), and the smart city and connected objects (11 tables). The analysis presented in this section was also enhanced by discussions in different satellite activities (input during classes; world cafés addressing the same themes without strictly following the method used during the co-construction tables).

To understand this section properly, we should note that the discussions addressed issues not only related to responsible AI development, but also those pertaining to data management (especially personal data and big data)—whether data from which algorithms learn, or data that, in some shape or form, is analyzed by AI. As these issues are interrelated, they were considered together for this analysis.

The scenarios served as launchpads for discussions during which two types of data were collected: perceptions of citizens regarding the **risks and issues** in AI development and **potential solutions** to address them (see scenarios, Part 1, Section 6, Annex 2). For the purposes of this section, the analysis remains descriptive and as verbatim as possible. The **main directions** expected in terms of responsible AI development refer to citizen recommendations that are not specified in concrete potential solutions. They nonetheless allow us to identify the main positions and standard expectations citizens have of AI development. When these expectations were debated during discussions or when citizens considered that responding to these expectations was an issue, they were considered in the issues category.

Each co-construction table was invited to choose two or three issues to be treated as priorities before 2025. Only issues that citizens considered priorities were taken into account in Section 3 for the purposes of this report. These priority issues were defined by citizens and classified, for each sector, according to the principles of the first version of the Declaration to which they refer. However, it is worth noting that just because certain issues were not considered priorities that they were not discussed, that they are less important, or that the principles

were not discussed for each sector. One single principle for each sector is detailed in this section.

A thematic analysis was made of all the discussions using NVivo software. The purpose of this analysis was to highlight citizens' perceptions of the scope of the risks and issues in AI development (see mind map, Table 3). These issues have been grouped into 12 categories, and are not mutually exclusive. We recognize that this is but one of many ways in which to classify the different discussions that took place. The potential solutions identified by citizens to address these issues were classifed into 11 main categories. These categories are mutually exclusive, thus allowing us to add quantitative data.

With regard to the quantitative data in this report, the number of times it occurs corresponds to the number of tables where each issue/potential solution was formulated through consensus, in keeping with the co-construction process. The total number of potential solutions (n=190) corresponds to those identified as priorities by citizens (since they were invited to clearly formulate them on posters). However, potential solutions that were mentioned during the discussions but did not explicitly appear on the posters were also taken into consideration.

Quotes from the report are presented so that they reference the co-construction table when they formulated by a group (consensus). Other quotes correspond to individual ideas formulated (written on Post-its by participants or copied verbatim by team members).

3. MAIN DIRECTIONS EXPECTED BY CITIZENS

Generally speaking, the participants recognized that AI had important potential benefits. Participants recognized the time savings that AI devices could bring particularly when it came to work and legal matters:

> "It would help reduce wait times to treat cases."
> — A participant

However, it was also mentioned that AI had to be developed with caution and from now on, to prevent harmful use although some consider the possibilities that AI opens up to still be limited. Introducing a framework was therefore recognized as necessary to prevent risks rather than trying to determine who is to blame when they occur:

> "You don't care so much about knowing who to sue when things go wrong, you want to find ways to make sure things don't go wrong in the first place."
> — A participant

The citizens highlighted the need to implement different mechanisms to ensure that quality, understandable, transparent and relevant information was being communicated. They also discussed the difficulty of guaranteeing truly enlightened consent.

Most participants recognized the need to align public and private interests to prevent monopolies from emerging, or limit the influence of corporations (which are sometimes seen as ungovernable) through more cohesive and legal measures. To the greatest extent possible, these mechanisms should be simple and evolve so they can adapt to the pace of AI development and maintain steady control of it. In the legal sector, certain participants mentioned a "divide" between technology (defined as quick, innovative, even abstract) and our institutions (often too rigid in their integration of technology) that are not able to deal with these changes in society. Some tables went as far as suggesting "nationalizing AI", which would then "become a public service, and programmers would be public servants". (Smart city and connected objects table, INM, Montréal, February 18, 2018, Connected refrigerator scenario.)

The participants also recommended ensuring that AI be considered in context, meaning different parameters must taken into account (e.g. mandatory or optional collection of data the algorithm learns from). These mechanisms should come from and involve independent, trained people to promote diversity and include those who are the most vulnerable, and protect different lifestyles.

Whatever the use, most participants insisted that AI must remain a tool, and that the final decision be made by a human being (whether a legal ruling, hiring decision or health diagnosis), which implies recognizing its limitations.

> "AI proposes, mankind disposes."
> — A participant

Protecting an individual's privacy and managing personal data were discussed in depth. For example, processing healthcare data should be managed in a unique way, given the highly sensitive nature of the information. It should therefore both promote control methods ranked according to type of use and adopt security as an operational mode. As for the workplace sector, participants recommended that employers be obligated to inform users of how their data is processed.

The participants were aware that these recommendations involve important institutional changes, and underlined the fact AI is not necessarily desirable to begin with.

"Just because you can, doesn't mean you should."
— A participant

The citizens generally agreed that impact of using AI in the different sectors—for both individuals and society as a whole—must clearly be measured to establish benchmarks without unduly hindering progress.

4. CITIZEN PERCEPTION OF RESPONSIBLE AI DEVELOPMENT ISSUES

4.1. INTRODUCTION

Citizens that took part in the co-construction days were invited to select two or three issues to address as priorities before the year 2025 with regard to responsible development of artificial intelligence.

Table 2: Priorities identified by citizens according to the principles of the Declaration (number of tables).

	Education	Legal system and predictive policing	Workplace	Healthcare	Smart city and connected objects	Total number of tables that consider these issues to be priorities
Responsibility	6	5	3	10	5	29
Autonomy	7	3	2	5	9	26
Privacy	6	5	1	9	4	25
Well-being	6	4	2	6	5	23
Knowledge	6	5	4	4	2	21
Justice	6	4	5	4	4	21
Democracy	1	4	3	1	7	16
Total number of co-construction tables	9	8	5	12	11	45

The responsibility principle was most often deemed a priority, followed by autonomy, privacy, well-being (individual and collective), knowledge and justice. It is worth noting, however, that they are all closely interrelated.

The principles of knowledge, responsibility, privacy, justice and democracy are presented below per sector. The autonomy principle, often selected as a priority, concerns preserving, even encouraging individual autonomy when faced with risks of technological determinism and reliance on tools.

It also raises the issue of the two sides to freedom of choice: being able to make your own choice when faced with a decision guided by AI, but also being able to choose not to use these tools without risking social exclusion. The freedom included in this autonomy principle regarding AI systems would involve any person's capacity for self-determination.

"Develop technologies that promote human autonomy and freedom of choice."
(Education table, Bibliothèque de Laval, March 24, 2018, Hyper-personalization of education scenario).

The well-being principle also holds an important place for participants. Participants at every table expressed a collective desire to move towards a society that is fair, equitable and promotes everyone's development. Well-being is therefore both a collective (touching on equity and accessibility issues within the justice principle) and an individual issue, aiming for everyone's fulfillment without hampering autonomy and privacy. Participants showed a preference for AI development "that would allow any individual to achieve personal and social fulfillment". (Education table, Bibliothèque Père Ambroise, Montréal, March 3, 2018, AlterEgo scenario.)

Broadly speaking, the well-being principle was also a call to maintain quality human and emotional relationships between experts and users in all fields.

MAIN ISSUES DISCUSSED PER SECTOR

EDUCATION

Six out of nine tables considered privacy, responsibility, well-being and knowledge issues priorities for the education sector. Discussions on issues related to the knowledge principle were especially relevant to broaching the subject of transforming human skill sets in the age of AI:

ISSUES RELATED TO THE KNOWLEDGE PRINCIPLE
(6 out of 9 tables)

For the theme of education, issues related to the knowledge principle concern changes in skill sets, given that the teaching profession and ways of developing and accessing knowledge are rapidly changing. This principle was mostly discussed from the perspective of how the learning relationship would change, how teachers' expertise would be challenged and how their work would have to change as a result. It was also mentioned in relation to the diversity principle: the need to cultivate a wide range of intelligences and relationships to knowledge.

"Redefining/transforming the nature of the relationship between teachers and students in the classroom and changing our relationship to knowledge."
(SAT Table, Montréal, March 13, 2018, Nao scenario).

"Human skills and abilities: the importance of developing many learning environments."
(Musée de la civilisation table, Québec City, April 6, 2018, AlterEgo scenario).

LEGAL SYSTEM AND PREDICTIVE POLICING

Five out of eight tables considered privacy, responsibility and knowledge issues priorities for the justice and predictive policing sector. Discussions on issues related to the responsibility principle allowed us to clarify the principle's scope:

ISSUES RELATED TO THE RESPONSIBILITY PRINCIPLE
(5 out of 8 tables)

The responsibility principle was formulated in two primary ways: as a demand for human accountability in legal rulings, and a concern for responsibility in decision-making (and any potential errors). From the citizens' point of view, the algorithm's lack of transparency goes against accountability, since it is difficult to know what factored into the decision. The responsibility principle is therefore linked to knowledge and transparency principles in that decisions should be explainable and preserve the skills and role of human beings in the legal system.

"[Justice] must remain a tool whose sole purpose is to protect individuals. Promote compassionate and equitable justice that accounts for idiosyncrasies and past experiences. Artificial intelligence must not have the right to judge human behaviour. The final decision must always require human intervention."
(SAT table, Montréal, March 13, 2018, Preventive arrest scenario).

"Transparency, accountability and responsibility when creating the tool, the data used, and the impact of this tool."
(SAT Table, Montréal, March 13, 2018, Parole scenario).

With regard to responsibility, citizens were concerned about overlooking human beings and human "agency". Failing to consider human dynamics and the ability for invidividuals to change shows a clear concern about a "static" vision of human beings provided by an algorithm, which would make its decisions problematic and unreliable. Participants were ready to make "agency" a principle of the Declaration in this workshop.

"We must take personal agency into consideration. The ability of each individual to change, to change their own course."

HEALTHCARE

Privacy and responsibility principles were considered priorities by 9 and 10 tables out of 12, respectively, in healthcare. Privacy issues were particularly significant for the sector given the relatively sensitive and invariably personal nature of healthcare data.

PRIVACY PRINCIPLE ISSUES
(9 out of 12 tables)

Participants identified different issues related to confidentiality and invasion of privacy. At issue was the possible invasion of privacy linked to developing and configuring AI systems (e.g. which should help avoid pirating, shortages and harmful use). Citizens also discussed "retroactivity" (use of data previously collected for another purpose) and accessing this data through private companies. In light of these issues, citizens' concerns included how to ensure that data isn't sold, and how to guarantee that the patient maintains control of their data (especially when it concerns private data), and holds full rights to it.

"To what extent are we willing to share our personal data (information) as individuals in order to feed healthcare services?"
(Musée de la civilisation table, Québec City, April 6, 2018, Digital twins scenario).

WORKPLACE

Issues on justice and knowledge were considered priorities for the workplace sector, (5 and 4 tables out 5, respectively). All tables that discussed AI development in the workplace, therefore, felt that issues concerning justice, equity and diversity should be addressed separately.

ISSUES RELATED TO THE JUSTICE PRINCIPLE
(5 tables out of 5)

Citizens had two primary concerns about the justice principle: ensuring an equitable sharing of AI benefits among all social groups and territories, and "including nondiscriminatory algorithms that favour diversity, inclusion and social justice". (Musée de la civilisation table, Québec City, April 6, 2018, AI as a compulsory step to employment scenario).

"Sharing AI benefits (productivity gains); equity among social groups, territories (cities and regions), taking vulnerabilities into consideration; the meaning of work in society and how it shapes our identities."
(Musée de la civilisation table, Québec City, April 6, 2018, Socially responsible restructuring scenario).

SMART CITY AND CONNECTED OBJECTS

For the smart city and connected objects sector, issues related to autonomy and democracy principles were considered priorities by 9 and 7 tables out of 11, respectively. Citizens felt that many issues could impact the democracy principle:

ISSUES RELATED TO THE DEMOCRACY PRINCIPLE
(7 out of 11 tables)

Participants discussed issues such as balancing collective interests and individual needs; managing access to public spaces and sharing said spaces, or even sharing the benefits from the development of AI technologies (particularly between individuals, the public sector and the private sector). They insisted on a need for and the difficulty of ensuring a collective (involving citizens) and enlightened (which implies a level of transparency in developing AI systems) decision-making process to define guidelines on connected objects. Citizens also questioned the true independence of public authorities in AI development, and discussed the risk of normalizing behaviour that could lead to marginalization, thereby possibly jeopardizing the democracy principle.

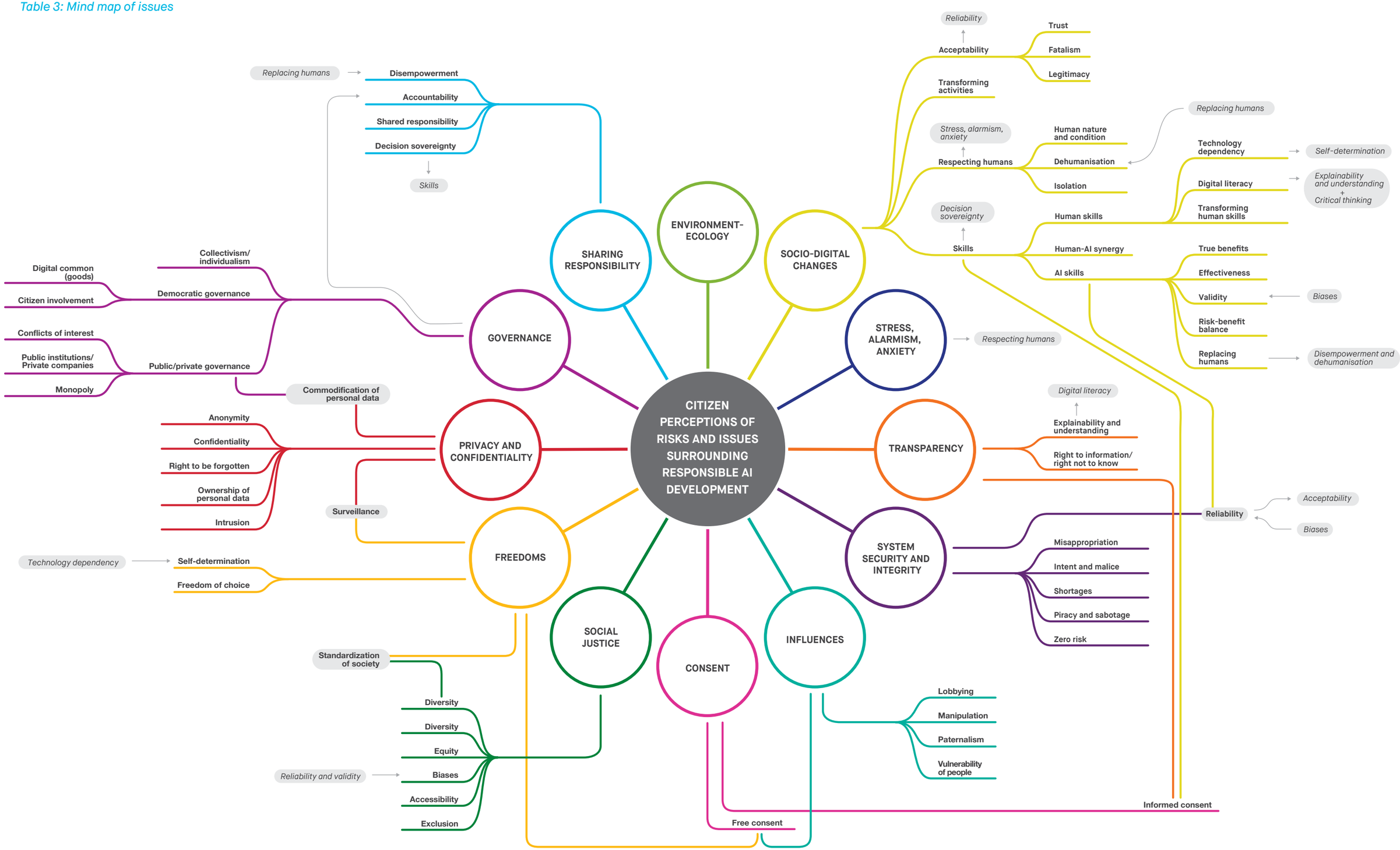
"How can we manage an intelligent transportation system democratically?"
(Bibliothèque du Boisé table, Montréal, March 17, 2018, Self-driving car scenario).

4.2.

MAJOR CATEGORIES OF RISKS AND ISSUES IN RESPONSIBLE AI DEVELOPMENT

Citizens identified 12 major risk categories and issues in responsible AI development during discussions of the different scenarios. These categories are not mutually exclusive, but offer a snapshot of various themes raised by citizens in responsible AI development and warrant special attention for the purposes of creating public policies. The following mind map presents the scope and diversity of the issues discussed, which have been classified into categories and subcategories. Sometimes, dilemmas or marked oppositions came out of the discussions. The following section provides a definition for each category, illustrated with examples taken verbatim.

Table 3: Mind map of issues



GOVERNANCE

COLLECTIVISM VERSUS INDIVIDUALISM

This category refers to a dilemma which pits protecting individual interests, choices or responsibilities against protecting collective interests, choices or responsibilities. The answer to this dilemma is an important issue that strongly depends on a normative position for which no consensus was reached.

"Ensuring that AI technology is a learning tool that serves the social and democratic ambitions of school as a public good."

(Education table, SAT, Montréal, March 13, 2018, Nao scenario).

"Digital twins: this is a very libertarian way of proceeding, which once again creates tension between individual and collective well-being."

"We are at a point in democratic life where the focus on the individual is so great that it will lead to a dictatorship."

"How can we ensure that self-driving cars maximize well-being? The sharing of public spaces? How can we reconcile the safety of the majority versus that of the individual?"

"Can public interests align with private personal interests and remain ethical?"

GOVERNANCE: PUBLIC VERSUS PRIVATE

Issues related how managing AI development would be divided between **public and private institutions,** and the inherent risks were also raised. These challenges were often presented as questions: How would this be shared equitably? Which of the two methods of governance is the most appropriate?

"Who is steering all of this? What powers will the organization or company hold over this tool? Will we be dependent on the company? If it becomes a national priority, what choices will be made for educational programs when it is implemented? Is it public? Private? The entire education ecosystem will be redefined."

More specifically, the **risks of conflicts of interests, commodification of personal data or the emergence of a monopoly were raised.** Participants particularly highlighted the risk of a conflict between private interests (essentially financial) and other interests, which could limit the independence of certain stakeholders or public institutions. The risk of commodifying personal data refers to issues related to the market value of data, the limitations of collecting data and the profits associated with it, particularly with respect to the protection of privacy. The emergence of a private monopoly in the governance of AI development was also a subject of concern.

"Avoid commercial use or interests that aren't educational when it comes to data collected and analyzed by AlterEgo"

(Education table, Bibliothèque Père-Ambroise, Montréal, March 3, 2018, AlterEgo scenario).

"How to avoid excessive commodification of data and people without their knowledge?"

(Smart city and connected objects, SAT, Montréal, March 13, 2018, Smart toy scenario).

"Excessive concentration of power (GAFAM), which prevents:

- Equitable sharing of AI benefits

- The arrival of new stakeholders (new business models, e.g. co-op)"

(Workplace table, SAT, Montréal, March 13, 2018, Socially responsible restructuring scenario).

DEMOCRATIC GOVERNANCE

Given that the discussion on governance often pits public institutions against private companies, issues on another alternative were raised: that of a participatory governance which involves citizens directly. These issues include the shared and collective management of open-access digital goods **(digital commons)** and the role of **citizen involvement** in current and upcoming governance (whether present or absent).

"Issue 3: Participatory democracy with a balance of power (states, social partners, businesses, unions, etc.)"

(Workplace table, Musée de la civilisation, Québec City, April 6, 2018, Socially responsible restructuring scenario).

Citizens recognized that the urgency of the situation and a certain technological determinism were factors that could harm participatory governance. The lack of time that would eliminate any possibility of a democratic process needs to be recognized.

"Urgency instead of taking the time to hold an informed and participatory democratic debate"

(Workplace table, SAT, Montréal, March 13, 2018, Socially responsible restructuring scenario).

SOCIAL JUSTICE

Citizens brought up different risks and issues regarding algorithms biases, access to AI and the consequent discrimination or exclusion of certain groups of individuals. They considered the impact of these risks on diversity and equity to be important issues.

"Implementing nondiscriminatory algorithms that foster diversity, inclusion and social justice"

(Workplace table , Musée de la civilisation, Québec City, April 6 2018, AI as a compulsory step to employment scenario).

Accessibility issues included how to guarantee access to AI and its uses. They are associated with restricting access of certain groups or social classes. Discussions were also held on the impartiality of algorithmic systems and their potential for discriminatory bias, namely data on which the algorithms are trained, as well as data collection or even the code itself.

"The values of justice (independence, impartiality, equity) prevail over technique when deploying these tools."

(Legal system and predictive policing table, SAT, Montréal, March 13, 2018, Parole scenario).

Citizens pointed out the discrimination that could arise if the first two categories of issues (accessibility and exclusion) are not adequately addressed: the discriminating effects of AI systems, whether by reinforcing existing discrimination

(e.g. gender or social status), or creating new discrimination (e.g. people who are not "connected"). **Discrimination** issues are closely tied to the risk of **exclusion** for some people, whether they voluntarily refuse to take part in the "digital society", or whether they are involuntarily excluded.

"What happens to people who don't have a digital profile? Are they at a disadvantage? Should we rely solely on AI for recruitment? Can AI truly grasp the hiring criteria? Do we have a choice if everyone else is doing it? And how do you evaluate a digital reputation?"

These risks led the participants to identify a protection issue for:

1. The **diversity** of intelligence, skills, individuals and society as a whole.

"Does AI simply reproduce the same intelligence that is taught in school? Wouldn't it be more beneficial to cultivate different types of intelligence?"

2. **Equity** so that AI operations led to decisions and recommendations.

"Sharing the benefits of AI (productivity gains). Equity between social groups, territories (cities and regions), taking vulnerabilities into account."

(Workplace table, Musée de la civilisation, Québec City, April 6, 2018, Socially responsible restructuring scenario).

FREEDOMS

This category refers to issues of maintaining individual freedoms, especially when it comes to freedom of choice—whether being able to make your own decision when faced with an AI-guided decision, or being able to choose not to use those tools without being socially excluded (which means that these issues are often closely tied to the previous category).

SELF-DETERMINATION

Citizens discussed the risk of algorithmic systems being overwhelmingly deterministic, particularly with regard to an individual's capacity for self-determination (as opposed to a risk of blind faith in technology).

"What concerns me the most is that the grandmother is excluded from the thought process. A robot nurse, fine, but what does the grandmother want? We have to ask people what they want."

FREEDOM OF CHOICE

Being able to make individual choices as well as the right to refuse to use technology or take part in a data collection system were also discussed.

"How can we ensure that an individual maintains their freedom to choose and doesn't become a slave to technology?"

"If we need everyone's data to create collective well-being, do we have to force everyone to share their data? And if some people refuse to do so, what impact will that have on the system? This is a societal choice that must be made."

STANDARDIZATION OF SOCIETY

The standardization of society addresses risk issues that arise when AI categorizes individuals for predictive purposes in healthcare, education, justice or mobility. This could lead to individuals being stigmatized and behaviours normalized instead of encouraging diversity.

"Risk of a standard profile (normalizing behaviours)"

(Smart city and connected objects table, INM, February 18, 2018, Connected refrigerator scenario).

SOCIO-DIGITAL CHANGES

This category refers to discussions and issues on social and societal changes that could result from AI development. These changes may (or may not) lead to a true "digital transition".

ACCEPTABILITY

Citizens repeatedly brought up the issues of acceptability and social buy-in when implementing AI. These discussions revolved around issues such as maintaining the public's trust in technology (AI) and in the different sectors that might use it. They also brought up issues of technological expectations and "technophobia". At times, there seemed to be a certain sense of fatalism, particularly toward technological determinism and a somewhat forced acceptance of AI development. The legitimacy of using AI in certain fields was sometimes questioned.

"Maintaining and promoting the population's trust in the justice system"

(Legal system and predictive policing table, Musée de la Civilisation, Québec City, April 6 2018, Parole scenario).

HUMAN SKILLS

Participants repeatedly discussed the impact of AI development on human skills. For example, they deliberated the transformation of human skills from the perspective of consequences (mainly negative) that AI development could have on knowledge and abilities.

"Fear of exceeding humans, human ability to be at 360° (whereas AI has excellent, very specific skills)."

"How can we ensure that dialogue with the patient is maintained (human contact) and that the doctor doesn't lose their expertise and independence?" (Healthcare table, Bibliothèque de Sainte-Julie, March 25, 2018, Intelligent hospital scenario).

A risk of dependence on technology (and more specifically, in this case, the use of AI) was brought up.

"We become dependent (on technology)"

"AI causes us become too specialized and takes us further away from general knowledge and independent learning."

The digital literacy issues refer to the need to educate the population on AI practices and issues, so people gain both the technical and critical skills required to function both as a worker and citizen in a digital society in transition.

"To guarantee that a device like AlterEgo is used intelligently, it is important that youth, parents and teachers be made aware of how the collected data is used. This raises a knowledge issue that entails an AI literacy approach."

AI SKILLS

Regarding AI skills, issues about the true advantages led to discussions questioning the potential benefits or uses of AI.

"How can we ensure that our AI tools respect the fundamental principles of our justice system?"

(Legal system and predictive policing table, Musée de la civilisation, Québec City, April 6, 2018, Parole scenario).

"Does AI fulfill its role of improving and providing access to the health and living standards of individuals/communities (rationalization, dehumanization of patient care, unexpected effects and actual efficiency of algorithms, etc.)?"

(Healthcare table, SAT, Montréal, March 13, 2018, Digital twins scenario).

Ensuring the efficiency and validity of AI, meaning the relevance of its use and skills, was also identified as an issue.

"We have to guarantee healthcare recommendations based on:

1. algorithms that are managed, validated, updated (based on scientific knowledge) and uncompromised (security/hacking);
2. complete, honest and unbiased data."

(Healthcare table, Benny Library, Montréal, March 18, 2018, Digital twins scenario).

"If AI draws wrong conclusions, how can we ensure that we are evaluating its performance? Inevitably, AI will evolve, and we will have to plan for mechanisms to validate the results and plan for continuous evaluation."

"Yes, after every decision there must be an evaluation of that decision. If we do not evaluate the performance and consequences of decisions made by the algorithm and we continue to use the algorithm, the AI will wind up basing itself on mistakes."

The risk of replacing humans was also brought up on many occasions, and was linked to the role of AI and the duties it could perform instead of a human, the advantages and inconveniences of its use as well as the way to share skills between humans and AI.

"AI will fill certain gaps in the education system, but is it the solution? Teachers' workloads will be considerably lightened, which gives them a break, but also raises the question of replacement."

More nuanced discussions highlighted issues of a balance between the benefits and the risks of AI and its skills, or the need to take these benefits and risks into account for responsible development.

"How can we implement AI into everyday objects while harmoniously developing society (cultural aspect, well-being, child development, candour) and living beings?"

(Smart city and connected objects table, SAT, Montréal, March 13, 2018, Smart toy scenario).

HUMAN-AI SYNERGY

This category refers to discussions about the advantages of human-AI synergy or the inconveniences of such a "collaboration". The main point of discussion was the synergy between the objectivity and systemization of AI on the one hand, and the subjectivity and empathic contextualization of humans on the other.

"Ensuring AI-teacher complementarity in terms of expertise and relationships with students."

(Education table, SAT, Montréal, March 13, 2018, AlterEgo scenario).

"How can we ensure that healthcare decisions aren't solely based on objective data but also consider the context and the user's choice?"

(Healthcare table, Bibliothèque du Boisé, Montréal, March 17, 2018, Healthcare insurance scenario).

"Objective justice from AI predictions versus subjective intelligence (based on experience)"

RESPECTING HUMANS

Respecting nature and the human condition were other issues raised by the citizens. These discussions led to questions of what defines a human being, what will be left of human beings, or how to put people first in the context of AI development and the importance it could take on.

"What is a human being? What are we keeping of human beings? What do we want to keep of human beings?"

The risk of the dehumanization of activities and services with AI development or the emergence of a new form of isolation—specifically caused by decreasing socialization, or delegating social relationships to robots—were also brought up repeatedly.

"The human aspect of care is lacking. The relationship between healthcare professionals and patients"

"How can we ensure human dignity and the place of human beings in the justice system?"

(Legal system and predictive policing table, Musée de la civilisation, Québec City, April 6, 2018, Parole scenario).

"This will also result in cases being standardized, and people themselves won't be sufficiently taken into consideration."

"Relationships with AI to the detriment of humans leads to growing solitude."

TRANSFORMING ACTIVITIES

This category refers to discussions surrounding societal changes that would come with AI development and the eventual digital transition in the various sectors concerned, at different levels (for example, AI transforms knowledge, the city, the conception of work, etc.)

"We're rationalizing health."

"Redefining/transforming the nature of teacher-student relationship in a learning environment and changing our relationship to knowledge"

(Education table, SAT, Montréal, March 13, 2018, Nao scenario).

"Will increasing mental capacity through transhumanism make education obsolete?"

"There's a risk of crystallizing law. The more decisions AI makes in a certain direction, the more likely it will be to rule the same way going forward."

"In 30 years, people will sleep, work, etc. in their car, which will cease to be a device used solely for transportation. Mobility will take on a whole new meaning."

PRIVACY AND CONFIDENTIALITY

ANONYMITY, CONFIDENTIALITY AND THE BENEVOLENCE DILEMMA

This category refers to respecting **anonymity** and **confidentiality issues**. Discussions were held around the real possibility of respecting anonymity with responsible AI development, how to ensure that "sensitive" data remains confidential, or how to restrict its access to certain people and uses that would be more justified than others. At times, AI was considered the problem; at other times, the solution to this type of issue. A **dilemma** surfaced on many occasions, especially in the field of healthcare. The dilemma highlighted the opposition between benevolence (which supposes collecting as much data as possible, and not just objectifiable data to ensure a more human and context-based approach by AI), and the respect of privacy and confidentiality (which would be challenged by this very data collection).

"Confidentiality no longer exists, it's a myth. We tried making data anonymous, it doesn't work. Now we can impose that only algorithms can see the data, not the human stakeholders that handle the data."

RIGHT TO BE FORGOTTEN

Discussions were also held on creating a right to be forgotten (being able to erase personal data), and the issues and impact of implementing it.

"Right to be forgotten (storage limitation), right to modification, right to suppression"

(Legal system and predictive policing table, Bibliothèque Père-Ambroise, Montréal, March 3, 2018, Preventive arrest scenario).

INTRUSION

Discussions about the risks of intrusion into people's private lives, breach of privacy and ways to guarantee protection were held on many occasions.

"How can we ensure that various components of private life (omission, property, consent, portability) are respected in the context of connected object use?"

(Smart city and connected objects table, SAT, Montréal, March 13, 2018, Smart toy scenario).

OWNERSHIP OF PERSONAL DATA

This category refers to issues related to ownership of personal data, its definition, the consequences of this ownership on privacy (to what extent does an individual own and remain the owner of their own data?) and the protection of people's "digital reputation".

"Data concerning private life should be the property of the people concerned and shared according to rules voted on democratically."

(Healthcare table, INM, Montréal, February 18, 2018, Digital twins scenario).

SURVEILLANCE

Issues of surveillance are linked to data accessibility and profiling, which raises concerns about (constant) mass surveillance of individuals that risks violating both privacy and individual liberties.

"How can we live healthy lives when we are constantly being watched?"

"Will we be able to track everyone's movements?"

"Could a higher power, government or company, take control of my vehicle?"

FREE AND INFORMED CONSENT

Discussions were also held on the capacity to consent to the use of AI and personal data.

FREE CONSENT

At issue was the true independence of individuals and their right to share (personal) data or not, to have a real impact on how it is managed or choose how it will be reused.

"Are we truly free to not share our data?"

"If we're sharing publicly, are we truly consenting to that information being reused?"

INFORMED CONSENT

The issue here ties into information mechanisms needed for individuals to consent in an informed manner; it concerns access to information and understanding this information. This issue is closely linked to citizens' digital literacy as well as transparency.

"The question of informed consent (for both students and parents) lies at the heart of issues of data collection and interpretation as well as student autonomy."

ENVIRONMENT/ECOLOGY

These issues concern the impact of responsible AI use and development on the environment, as well as its energy costs.

"We don't often talk about the environmental aspect: storing data, stockpiling outrageous amounts of data and the inherent energy costs."

INFLUENCES

These issues refer to concerns about AI's influence (whether undue or not), or potential for manipulation. To maintain a certain freedom in the choices guided by AI and avoid placing blind trust in these devices, citizens recognized the need to cultivate critical thinking among individuals who use AI.

LOBBYING

Citizens worried about AI creating a new type of lobbyism, which could yield too much power and influence over the healthcare system, connected objects or self-driving vehicles.

"Should it be up to politics to determine which algorithm will be used? What about a lobby for algorithm designers?"

MANIPULATION

Participants worried about the risk of users being manipulated as actions and decisions become increasingly influenced by AI mechanisms, whether unknowingly or through more explicit incentives.

"To what extent can a machine influence our decisions? Do we know what impact a connected refrigerator's suggestions will have on our daily lives?"

"Insidious influence on our behaviours without us asking for it or accepting it"

"Influence risks: How can we make the risk of influences (consumption, judgment) linked to connected object use visible? How can we ensure everyone's interests (consumers, citizens, companies) are respected? Who determines the guidelines for developing these (eco) systems, and how do they go about it?"

(Smart city and connected objects table, SAT, Montréal, March 13, 2018, Smart toy scenario).

PATERNALISM

Exposure to various forms of paternalism and control (from companies, the State) was mentioned on more than one occasion. It could be increased through incentive systems, but also through the depersonalization of relationships (namely patient-heathcare provider relationships).

VULNERABILITY

Citizens recognized that not everyone is as vulnerable to the influence risks presented. Special protection of those who are most vulnerable was highlighted as an important issue.

SHARING RESPONSIBILITY

This category refers to issues of shared responsibility in responsible AI development and the consequences of decision-making.

DISEMPOWERMENT

Disempowerment here refers to concerns about the risk of disempowerment in AI development, which could translate into delegating this responsibility to algorithms (considering their growing autonomy or the perception of a growing autonomy).

"Risk of disempowering the teacher who would defer to 'diagnosis syndrome', combined with the risk of reinforcing a certain student profile."

(Education table, Bibliothèque Père-Ambroise, Montréal, March 3, 2018, AlterEgo scenario).

"It creates a lack of accountability: say I'm hyperactive, the machine confirms it, so I put in less effort. But you have to be part of the solution, buddy. The way of working will change. A teacher's duties are going to change, that's for sure."

"Knowledge is tied to responsibility. There's a risk of disempowerment if there is a loss of knowledge. A loss of critical thinking from judges and other people."

"How can we ensure that AI remains a service and that the various stakeholders (individuals, programmers, society, etc.) aren't disempowered, remain vigilant, and that individuals are always in control?"

(Smart city and connected objects table, Musée de la civilisation, Québec City, April 6, 2018, Connected refrigerator and Carbon footprint scenarios).

ACCOUNTABILITY

This issue is about identifying who is responsible or accountable in various situations concerning AI development (the user, the developer, the algorithm, etc.?).

"Who holds the learning data, who uses it, for how long? Who is protecting it?"

(Education table, Bibliothèque de Sainte-Julie, March 25, 2018, Nao scenario).

"Who is steering all of this? What power does the organization or company hold over this tool? Will we depend on this company? If it becomes a national priority, what choices will be made for educational programs when it is implemented? Is it public? Private? The entire educational ecosystem will be redefined."

"Who manages the algorithm, who controls it, who supervises the person programming it?"

SHARED RESPONSIBILITY

Discussions were also held on sharing responsibility in AI development, the complexity of this sharing and the need to take all responsibilities and stakeholders into consideration.

"The issue of the individual and shared responsibilities, which may be conflicting, of various stakeholders (governments, healthcare professionals, patients, private companies, researchers and managers, etc.)."

(Healthcare table, SAT, Montréal, March 13, 2018, Vigilo scenario).

"Issue 2: Define everyone's roles and responsibilities (institutions, students, teachers) to provide a framework for implementing AI"

(Education table, SAT, Montréal, March 13, 2018, Nao scenario).

"I don't know of any teachers that shirk their responsibilities towards their students. But we need to involve as many people as possible, adopt a multidisciplinary approach. Not make the teacher the sole person responsible for AI or AI diagnoses. Ensure that using AI for educational purposes is a shared responsibility."

DECISION SOVEREIGNTY

Issues of decision sovereignty echo the normative expectations detailed in the recommendations ("Main anticipated directions") which state that AI must remain a tool, an assistant or an additional information resource. These recommendations were made following discussions on issues of decision sovereignty; that is, whether humans or AI should have the last word.

"Algorithms should always give advice, not make decisions. The absence of human moderation is problematic, as algorithms don't take all aspects of an individual into consideration."

(Healthcare, Bibliothèque Père Ambroise, Montréal, March 3, 2018, Digital twins scenario).

"The problem with interpreting Alterego's diagnosis is that we can't forget that human intervention is necessary. We can't rely solely on a machine."

"We delegate a lot of micro-decisions to AI and interconnected systems, at the expense of humans."

STRESS—ALARMISM—ANXIETY

Participants worry that AI development will induce stress, alarm or anxiety due to information and notification overload or a lack of human contact, among others.

"How will students develop academic independence and learn to manage their stress and emotions when they no longer have access to AlterEgo during their post-secondary education?"

(Education table, Benny Library, Montréal, March 18, 2018, AlterEgo scenario).

"We must guarantee an individual's well-being when informing and treating them: not be alarmist."

(Healthcare table, Benny Library, Montréal, March 18, 2018, Digital twins scenario).

SYSTEM SECURITY AND INTEGRITY

AI and data system **reliability** issues were discussed at several levels: validity, infallibility and robustness, the integrity of systems and the people managing them. System vulnerability (bugs, errors, etc.) and impact of breaches on different reliability parameters were also raised. The risk of system **outages** and managing these risks were also among the issues brought up. These issues are closely linked to AI skills and biases. Citizens worried about risks of **piracy or sabotage** of algorithms and collected data, whether or not it was intentional, and the risks associated with potential **misuse** of data and algorithms (without necessarily amounting to piracy) and the problems it could cause.

"I don't want to be judged later on for things I did in the past."

"What if a hacker took control of the educational development of certain students? Or if parents could have an even greater impact on their children's grades? The hacker or the parents could choose the content, and therefore how AlterEgo interprets the data. For example, parents who do not want their child pursue a career in the arts could use AlterEgo to these ends."

Intent and malice in problematic or unsecured use of AI were identified as important parameters. Citizens pointed out that it was difficult to differentiate a malicious act from a problematic act that had good intentions, and the consequences of this distinction.

"Even with good intentions, we can cause problems (inaccurate model)."

"How can we distinguish temporary behaviour with no harmful intent from a genuine decision to carry out a crime?"

On many occasions, discussions revolved around a **zero risk** possibility, and whether it was truly desirable.

"Should the zero accident objective be reached at all costs? Is this objective really worth it?"

A number of dilemmas were identified during discussions on protection of security:

- Transparency (guaranteeing transparency could increase risks of piracy)
- Efficiency (ensuring the greatest possible security involves a compromise with system efficiency, as it must be secure without becoming inoperative)
- Respect of privacy and individual freedoms (in the specific case of preventive arrests, which impose surveillance in the name of public safety)

TRANSPARENCY

The issue of transparency was formulated as the ability to understand an algorithmic decision and to react to it, whether as an ordinary citizen or a professional using AI for their job.

EXPLAINABILITY AND UNDERSTANDING

These issues pertain to the explainability of a decision and the "black box", the importance of showing the process that leads AI to a result or the intelligibility of information and the importance of it being explainable.

"Transparency of the variables used, data, parameters. Explaining a decision in plain language."

(Workplace table, Mordecai-Richler Public Library, Montréal, March 10, 2018, AI as compulsory step to employment scenario).

"The complexity of the world of algorithms does not allow us to understand how AI proceeded (...) We don't require that much transparency from judges, so why should we request as much from the algorithm?"

RIGHT TO INFORMATION VERSUS RIGHT NOT TO KNOW.

This dilemma surfaced particularly in the healthcare sector and sets the right not to know (the entire range of diagnostic predictions provided by AI, for example) against to the right to know (to respect a patient's autonomy and consent). The right not to know could be justified in the name of benevolence (if certain recommendations are alarmist and uncertain).

5. POTENTIAL SOLUTIONS AND FRAMEWORK FOR RESPONSIBLE AI DEVELOPMENT

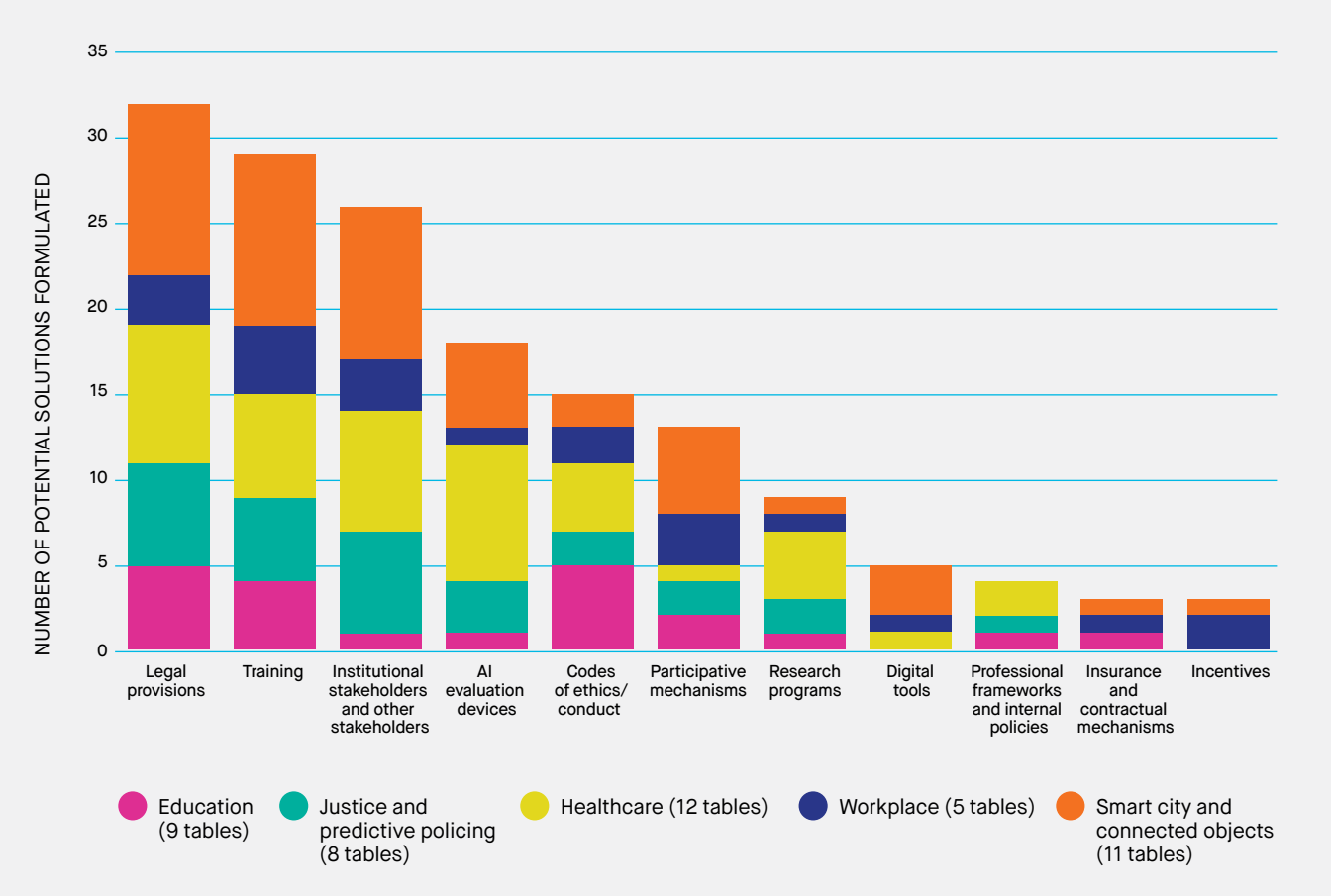
5.1.

INTRODUCTION

Citizens who took part in the co-construction days were invited to propose solutions to the previously identified issues. A total of 190 potential solutions were formulated and adopted by consensus during these activities (although other suggestions may have been discussed during the tables). By potential solutions, we mean concrete mechanisms that citizens put forward to respond to the previously identified issues.

Only possible solutions written on posters were counted. However, other recommendations were discussed or suggested (during the drafting of headlines and leads or in discussions). For the sake of coherency and feasibility, they were not included in the total number of recommendations, but were considered and analyzed when writing this section.

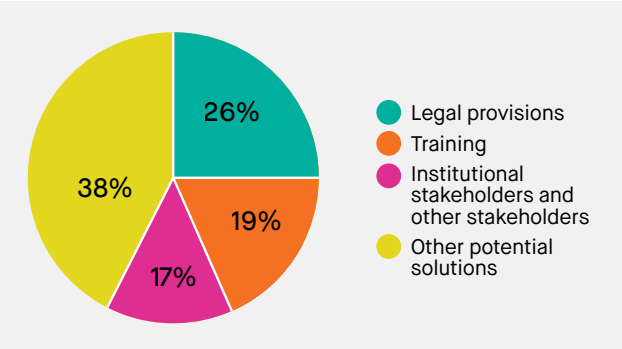
Table 1: Potential solutions proposed to respond to the issues identified



All co-construction tables agreed on three (3) key potential solutions to guarantee socially responsible AI development, regardless of sector:

1. Legal provisions
2. Training for everyone
3. Identifying independent key actors for AI management

Table 4: Three key potential solutions at all tables



Regardless of the sector, all tables agreed on recommending that a legal framework adapted to the reality of AI development and personal data management (especially big data) be implemented. These restrictive provisions all refer to rights or laws. They could be laws and regulations, defending new fundamental rights, or even public policies (ranging from implementing social programs and a charter to creating digital citizenship).

Implementing training that was accessible to all was also strongly recommended, both for professionals in the affected sectors (to ensure adequate use of AI systems in their work) and the general population (to ensure everyone can participate in the debate and acquire basic digital literacy).

Citizens also identified the institutional stakeholders and the key independent and competent stakeholders (existing or to be created) who would oversee responsible AI development. The stakeholders identified are people (e.g. ombudsman, auditor, life and well-being commissioner) or groups of people (e.g. setting up an artificial intelligence centre for civilian security, a 1–800 number against discrimination by connected objects or a ministry of data ethics and digital protection).

By recommending these three main mechanisms as potential solutions, a distinct trend emerges in the position held by Quebec citizens who took part in AI governance activities: it should primarily be handled by the State. Indeed, implementing incentives for businesses, or insurance and contractual mechanisms that correspond to a more liberal management were the least recommended potential solutions. These recommendations are nonetheless coherent and instructive. Citizens at different tables agreed on developing incentives—to encourage responsible development—and implementing diversity quotas (which reward companies that guarantee not to exclude or discriminate against certain minorities through AI biases) or funding for companies that help employees transition when jobs are being replaced by AI. Creating contracts between the various AI development stakeholders and its users, or insurance mechanisms to guarantee the protection of individuals in the advent of AI development was also suggested.

In all sectors, citizens suggested creating technical and ethical evaluation mechanisms for AI. Establishing a certification (or label) system as an ethical guarantee was suggested on many occasions, in particular. Different tables also recommended implementing a code of ethics (whether updating the existing code or creating new ones) and participatory mechanisms (e.g. co-constructions or public consultations) to guarantee that AI development and management remained democratic. Establishing professional frameworks (and different internal procedures for companies and institutions) that were not codes of ethics were also discussed.

The importance of implementing research programs in various disciplines (e.g. philosophy, social sciences, bioethics) to cultivate new knowledge and create digital tools (e.g. digital and interactive healthcare consent forms, personal digital file in the workplace sector) was also raised.

The following sections present the potential solutions formulated by citizens per fields of AI application. These potential solutions, defined

through concrete mechanisms, were not all discussed and developed to the same degree. Although it is evident that it is hard to imagine implementing all these recommendations given their diverse and somewhat contradictory character, an comprehensive presentation does offer, however, an especially robust global vision of the variety of solutions considered by citizens in AI management.

5.2.

EDUCATION

Table 5: Potential solutions or general guidelines for the education sector

	Number of potential solutions formulated
Legal provisions	8
Training	7
Codes of ethics/conduct	5
Participative mechanisms	2
Institutional stakeholders and other stakeholders	1
AI evaluation devices	1
Research programs	1
Professional frameworks and internal policies	1
Insurance and contractual mechanisms	1
Total	27

LEGAL PROVISIONS

Participants raised the need for creating and tightening certain laws in AI development in education. For example, a right to be forgotten was recommended regarding data use, as was an "expiration date", and no default sharing with other services unless there is a serious reason to do so. The right to be forgotten was often identified as the need to create a "data destruction policy" to allow students to reshape their identities and grow as individuals. The need to reinforce protection of privacy (particularly when it comes to data from youth) and transparency concerning data collection (namely by encouraging formats that are easily understood by users) was also brought up. For participants, a legal framework where "under no circumstance should the use of artificial intelligence limit a user's future possibilities, whether social, economic, etc." (INM table, Montréal, February 18, 2018, AlterEgo scenario), should be implemented.

Other initiatives were also formulated, such as creating a rule so that parents and students can choose to use AI devices or not, defining industry involvement in the education system to ensure ethical use of AI, and finally planning for strategies (through public policies) that would avoid "education hacking" by keeping data encrypted.

Furthermore, some citizens suggested creating a law or regulation that aims to "develop a common language (inspired by healthcare with food nutrition labels on processed foods) to bridge the gap between technology and its users" (Laval Library table, April 24, 2018, Nao scenario).

TRAINING

With regard to education, participants recognized the need to be proactive in implementing training for the entire community affected by AI development in this sector. Training would cover digital literacy, media literacy, as well as ethics and issues related to integrating AI in an educational environment. This could be, for example, digital literacy training for both parents and students, or it could be directly integrated into initial citizen training.

Citizens also recommended training education professionals specifically, for instance by including the development of work skills "in tandem" with AI devices in the curriculum for the basic and university training of teachers (e.g. a certification for the B.Sc. or an accreditation system). This training would be technology-based (how to use AI), but also geared towards teaching techniques with AI (how to organize teaching plans and emphasize that knowledgeable professionals direct AI, not the other way around).

"Accrediting change agents (both psychoeducators and active teachers) per teaching establishment to gradually integrate AI in an academic environment."

(SAT table, Montréal, March 13, 2018, AlterEgo scenario).

The importance of establishing adequate training was also raised. The training's purpose would be to provide stakeholders with appropriate information to assume responsibility for AI, and to discourage teachers from putting blind faith in educational AI devices. This training would help accelerate stakeholders' understanding in learning environments and mobilize them to develop AI so that learners became autonomous and equipped to deal with these realities. This training will help develop human skill sets and provide impetus to guide and even redefine future AI development.

"Raise awareness around responsible use of AI and promote a diversity of relationships to knowledge."

(SAT table, Montréal, March 13, 2018, Nao scenario).

CODES OF ETHICS/CONDUCT

Citizens also recommended implementing codes of professional conduct or ethics for teachers, which would focus on different ethical principles (e.g. justice) for AI use in an educational environment. These codes would provide a professional framework to prevent teachers from becoming disengaged as well as the risks of harmful use, profiling or discrimination.

"Ensure that AI use is a shared educational responsibility (support staff, family, teachers, robot)"

(SAT table, Montréal, March 13, 2018, AlterEgo scenario).

"Teaching while preserving the relational and emotional quality of human interaction."

(Bibliothèque de Sainte-Julie table, March 25, 2018, Nao scenario).

PARTICIPATORY MECHANISMS

Citizens suggested establishing open-source AI communities in public libraries to crack the AI "black box". The idea of leading general assemblies through consultations on socially responsible development of AI in education was also suggested.

"Consultation in the field of education to assess the current situation and define the roles and responsibilities of each player"

(SAT table, Montréal, March 13, 2018, Nao scenario).

INSTITUTIONAL STAKEHOLDERS AND OTHER STAKEHOLDERS

Citizens suggested creating a permanent Quebec multi-stakeholder committee that would be made up not only of department officials, but also representatives for parents, students, teachers, librarians and researchers. This would be a space for public debate and would serve as a counterbalance to private companies. This committee's mandate would be to advise the government (binding recommendations); prepare codes of ethics and training; introduce and oversee open source licences and consult with citizens. Citizens also recommended setting up ethics committees that would conduct consultation processes at every level of a technology's evolution, while ensuring its social acceptability. The idea of creating a joint, inclusive and diversified committee made up of educational stakeholders was also suggested. Citizens felt that a department should be responsible for creating this committee. Lastly, certain participants recommended creating a "Department of technological access and integration for training and certifications" (Bibliothèque de Laval, March 24, 2018, Nao scenario)

AI EVALUATION MECHANISMS

Participants felt that creating certifications was mandatory, particularly to ensure that certain standards were upheld, such as respect, conscious choice and freedom. Also, some certifications could guarantee that algorithms would not be used to replace teachers. Participants recommended tests and classroom observations to ensure this type of tool does not impede students.

RESEARCH PROGRAMS

Citizens recommended the joint or parallel development of technology and human creativity through research programs led by interdisciplinary stakeholders. These programs could focus, for example, on technology and mental health, ensure freedom of choice in using AI and safeguard human autonomy in decision-making. They also recognized the need for AI in for educational research, to intervene as early as possible in a child's learning.

PROFESSIONAL FRAMEWORKS AND INTERNAL POLICIES

Citizens believe that schools that integrate AI should do so responsibly. To this end, they recommended two potential solutions: implement incentives that encourage "schools to adopt internal policies to provide a framework for AI integration" (SAT table, Montréal, March 13, 2018, Nao scenario) or establish protocols or guides that help identify certain benchmarks to help integrate AI responsibly in schools.

INSURANCE AND CONTRACTUAL MECHANISMS

Citizens stated there must be a clear commitment to preserving the well-being of students. This commitment could be a "moral or social contract" that would have to be signed by all stakeholders. Implementing it would help "clarify the degree of responsibility in protecting student well-being" (Musée de la civilisation table, Québec City, April 7, 2018, AlterEgo scenario), but also provide teachers with the right to opt out.

5.3.

LEGAL SYSTEM AND PREDICTIVE POLICING

Table 6: Potential solutions or general guidelines for the legal system and predictive policing sector

	Number of potential solutions formulated
Legal provisions	9
Institutional stakeholders and other stakeholders	7
AI evaluation mechanisms	5
Training	5
Codes of ethics/conduct	2
Participative mechanisms	2
Research programs	2
Professional frameworks and internal policies	1
Insurance and contractual mechanisms	1
Total	34

LEGAL PROVISIONS

With regard to the legal system and predictive policing, laws and regulations on transparency must be established: private and public companies that collect criminal data must be transparent, and the decision-making processes by algorithms must be able to be explained and interpreted. Explaining the decision must come with measures that allow access to mobilized algorithms and ensure they are explainable and intelligible. As an initial transparency mechanism, many deliberation tables suggested that the AI used in the legal sector—even all public sector AI—be developed in open code, under free licence. From a legal standpoint, it's about guaranteeing "the right to full answer and defence", in particular, being able to challenge a decision by raising procedural or formal defects (Musée de la civilisation table, Québec City, April 6, 2018, Parole scenario).

This call for transparency goes hand in hand with establishing legal provisions that allow for what is considered a fundamental right to be judged by a human being to preserve procedural justice and individual sentencing, but also that the appeal process for a computer-based decision is always overseen by a human judge. Many debates revolved around conciliating human and artificial stakeholders

in this process, underlining the need for the law to adapt to a new technological reality that included AI in legal decision-making. The consensus was as follows:

> "The right to appeal before a human judge: The appeal procedure for a computer-assisted decision must always be heard by a human judge."

(Musée de la civilisation table, Québec City, April 6, 2018, Parole scenario).

In the scenario for using AI for preventive policing, citizens expressed the desire to establish a "framework that allows us to go beyond and eliminate biases, discrimination and abuse of power" (SAT Table, Montréal, March 13, 2018, Predictive arrest scenario) and tighten laws around consent to ensure it is truly enlightened. They also put forward the idea of limiting public and private stakeholders' access to private data, such as "private conversations on digital platforms" (Bibliothèque du Boisé table, March 17, 2018, Preventive arrest scenario) and enforcing a "right to be forgotten, to modify and correct data as well as a right to personal access to the data collected" (Bibliothèque Père Ambroise table, March 3, 2018, Preventive arrest scenario).

CODES OF ETHICS/CONDUCT

Citizens recommended establishing a declaration of principles, a code of ethics or conduct within companies, for the various professional bodies concerned or all individuals with access to algorithms. These codes would deal with consent, confidentiality, neutrality and how to protect human diversity. They would namely mitigate the speed with which AI technologies are developed, and the possibly ungovernable character of the companies that commercialize them.

> "Put the declaration of principles first: Live together harmoniously," meaning that we should "continuously review and optimize algorithms so they always serve humanity and human diversity".

(Bibliothèque Père Ambroise table, March 3, 2018, Preventive arrest scenario).

TRAINING

Participants highlighted the need for awareness campaigns to develop citizens' critical thinking on AI, their right to privacy and the sharing of their data. Learning should also include digital literacy and basic skills that must be developed in primary school. The training should ensure that citizens are aware of the programs and types of data used, that they have the knowledge and necessary tools to make educated choices and better manage the information they are sharing (e.g. as an information campaign, a public event or a discussion).

Certain tables also recommended introducing mandatory training for all high school students:

> "The training would include three steps:
> 1. The essence of AI
> 2. Functions and roles of AI
> 3. Ethical responsibility of AI"

(INM table, Montréal, February 18, 2018, Preventive arrest scenario).

Citizens also raised the need for training professionals in the field. Namely, by recommending that the judicial council define the type of training and adopt regulations to educate judges on new technological realities, so that they understand how AI works, the ethical issues related to AI and the impact of algorithmic decision-making on individuals and professionals.

PARTICIPATORY MECHANISMS

Citizens brought up the need to hold a major public consultation prior to using AI in a legal context and implementing any type of framework. The theme "For or against AI in law" would be at its core. The goal of the consultation would be to establish specific conditions for AI development in the sector prior to implementing legal AI applications. The consultation should be ongoing and evolve with new developments.

Citizens also suggested implementing consensus-based decision-making mechanisms that could be a co-construction session involving all stakeholders (professional bodies, associations, litigants, Department of Justice, industrial sector, etc.) when AI tools are acquired and deployed. They also highlighted the need to include AI users in this sector (e.g. judges, lawyers), who must be involved when selecting the product. In short, citizens felt that there was a need for consensus-based decision-making with stakeholders during the acquisition and deployment of the tool.

RESEARCH PROGRAMS

Citizens also recommended implementing university, industry and multidisciplinary research centres or programs focusing on the social, ethical, economic and political impact of AI on our society and the lives of individuals. Participants felt it was crucial to:

"Ensure that research generates solid data about the use of AI in law."

(SAT table, Montréal, March 13, 2018, Predictive justice scenario).

INSTITUTIONAL STAKEHOLDERS AND OTHER STAKEHOLDERS

As they pondered how to adapt AI tools to respect the fundamental principles of the legal system, many participants raised the need to create an independent organization to certify AI tools. It would not be to certify the tool's decision, but rather the algorithm's decision-making process. This would help ensure that the data is free of bias and that the algorithm is transparent and interpretable. Monitoring the tool's quality should continue after certification, through an audit process, for example. Many tables suggested that these independent organizations be hybrid entities (made up of public/private stakeholders, engineers, law professionals, social science researchers, ethics philosophers, etc.).

"The purpose of this entity would be to control AI. It would identify potential biases and would be achieved through co-construction"

(SAT table, Montréal, March 13, 2018, Predictive justice scenario).

Participants also brought up the need to create an independent group or body—made up of citizens and members of society—as a recourse in the event that certain principles of fundamental rights or justice were not respected. Likewise, they suggested creating a department of data ethics and digital protection, especially to preserve diversity and live in harmony with others.

Lastly, some participants suggested creating an "Artificial intelligence centre for civilian security" (AICCS) to ensure freedom, security and justice for all. "This centre, made up of citizens and professionals" aims to control "The abusive use of AI and highlights its first role and ultimate purpose, which is to be a tool that serves citizens." (INM table, Montréal, February 18, 2018, Preventive arrest scenario).

AI EVALUATION MECHANISMS

Citizens regularly put forward the need for institutional stakeholders to create standards and introduce certifications (on the creation and training processes for algorithms) that aim to protect rights and freedoms in the age of AI. They also talked about leading multidisciplinary studies a priori and impact studies a posteriori, running tests and reviewing and updating algorithms. Some also suggested creating a certification for "clear data and explicit intentions"

(Bibliothèque Père-Ambroise table, Montréal, March 3, 2018, Preventive arrest scenario). This would be an ethical certification on data dissemination and its objectives for the corporate world and government departments, in particular.

PROFESSIONAL FRAMEWORKS AND INTERNAL POLICIES

Participants expressed concerns that companies commercializing AI would become extremely adept at avoiding any form of control. They had two recommendations to this end. First, implement an ethical procedure within companies. Second, oblige private or public companies to write a mandatory annual report on significant incidents linked to AI use, out of a concern for transparency.

INSURANCE AND CONTRACTUAL MECHANISMS

Participants expressed the need for trade secrets to be lifted for legal stakeholders and citizens. This could be accomplished by introducing contracts between industry and legal stakeholders that specified the need to make the code open, examinable and verifiable for legal stakeholders and citizens.

> "AI code should be open-source and the decision should be as explainable as possible"

(SAT table, Montréal, March 13, 2018, Predictive justice scenario).

5.4.

WORKPLACE

Table 7: Potential solutions or general guidelines for the workplace sector

	Number of potential solutions formulated
Training	8
Institutional stakeholders and other stakeholders	5
Legal provisions	7
Incentives	3
Participative mechanisms	3
Codes of ethics/conduct	2
Digital tools	1
Public policies and guidelines	1
AI evaluation mechanisms	1
Research programs	1
Total	32

TRAINING

Citizens recommended implementing workplace training for everyone so that knowledge on current AI development issues could be shared. This training would help reinforce digital literacy and individual skills, as well as guarantee that citizens and future generations are aware, trained and ready for the current digital transition.

This workplace training would need to take rapid changes and uncertainties in AI development into account. This could be achieved by upgrading school curriculums, establishing awareness or support programs by the government (e.g. digital literacy programs for adults) or ongoing training for professionals. In particular, citizens came up with the idea of government agencies establishing public training for AI and digital realities to so that every segment of the population could benefit from its development.

"A major awareness program on the transition to AI as well as support programs launched by the government"

(Musée de la civilisation table, Québec City, April 6, 2018, Responsible restructuring scenario).

To avoid challenges related to AI use in recruitment, human resources professionals should also follow rigorous training on the methodological foundations of algorithms, digital data collection and the legal framework, and biases that are present or possible in AI analysis. An accelerated upgrading process and professional programs must be created with CEGEPs, universities, government departments, and professional bodies impacted by AI (e.g. law, healthcare).

INSTITUTIONAL STAKEHOLDERS AND OTHER STAKEHOLDERS

Citizens suggested creating three types of institutional stakeholders: a Crown corporation for AI in Quebec, an interdepartmental committee that advises the premier and governance committees in all companies that use AI in their recruitment process.

The mandate of the Crown Corporation for AI in Quebec, or NSAIQ (National society for artificial intelligence in Quebec) would be to support the digital transition through public policy expertise and provide assistance to private and public organizations, while opening up a democratic dialogue for AI implementation in public services:

"Its different mandates are:

> Ensuring AI expertise for drafting public policies (work, jobs, training, land use planning, education, etc.)

> Organizing democratic testing and implementation of AI in society and public services

> Supporting public and private companies throughout the transition

> Supporting and advising ministers on social programs in Quebec

> Help Quebec in international work groups"

(Musée de la civilisation table, Québec City, April 6, 2018, Socially responsible restructuring scenario)

The suggested multiparty committee would be a permanent, joint committee on economy, jobs, education and culture (inspired by the Digital Strategy). It would act as a direct advisor to the premier. This committee would allow the government to benefit from expertise independent of consultants and would not rely on private companies or third parties.

To ensure best practices for companies in AI-assisted recruitment, the suggested governance committees would be established in every company that uses AI in its recruitment processes. The mandate of these committees (one per company) would be to ensure that the code of ethics for human resources advisors is respected (see "code of ethics"). It would also ensure ongoing training for recruiters to ensure that they remain watchful for unpredictable biases that can occur at any time, and take into account the evolving nature of AI. Each company's committee would be multidisciplinary, made up of AI experts, HR experts, and people working outside the fields of AI and HR to allow for diverse opinions and experiences, and maintain a certain independence. Implementing an AI office in companies was also suggested to allow workers to see if AI use by an employer is acceptable from a legal standpoint.

LEGAL PROVISIONS

The legal provisions suggested by participants sought to address two main issues: guaranteeing human-focused AI development with an update to the Charter of Human Rights and Freedoms, and protecting (and reviewing) personal data.

"Updating the Charter of Human Rights and Freedoms to include AI and put humans first."

(Musée de la civilisation table, Québec City, April 6, 2018, AI as a compulsory step to employment scenario)

In a legal context, the idea of company accountability was defended, particularly when it came to protection of privacy: in the event of a predictive model likely coming into conflict with the existing legal framework, the company responsible for the model should communicate the information needed to evaluate its impact. Similar to the protection of privacy, the protection of personal data at work could be ensured by a regulation requiring that users be made aware that their data is being processed, as well as what data the company has, who has access to it, for what purposes, since when and for how long. All individuals should be able to access and understand this information, which could be stored in a personal digital file (see "digital tools").

Moreover, regarding the risk of exclusion inherent to holding compromising data, participants suggest allowing a form of "digital rehabilitation" for citizens who may be unfairly judged by digital footprints. A legal framework should be drafted to guide this kind of right to be forgotten, particularly to deal with the delays and the specific nature of this digital rehabilitation. This would also allow citizens to choose what information about them is made available, namely on social media.

"We must respect the existing legal framework, especially fundamental rights that already prevent discrimination when hiring. We suggest adding the right to digital rehabilitation (or the right to be forgotten) [so people aren't unjustly sidelined for digital footprints consulted by potential employers]."

(SAT table, Montréal, March 13, 2018, AI as a compulsory step to employment scenario).

Citizens also discussed creating anti-discrimination laws for algorithms or a minimum guaranteed income to help protect jobs lost in the transition.

Participants also stressed that the law needed to adapt to the many issues compounded by AI, but remain somewhat flexible in its review process to respond to the evolution of AI and its effects. Participants also recommended an "experimental approach" to avoid introducing regulations that are destined to change quickly.

INCENTIVES

Citoyens recognized the need to implement various incentives to encourage responsible AI development in the workplace, particularly with respect to the digital transition and protecting employee well-being. They brought up the need to reconsider how society directs public funds to AI and to demand socially responsible investments.

"Directing investments towards responsible AI for the common good."

(SAT table, Montréal, March 13, 2018, Socially responsible restructuring scenario).

These investments, along with employee pension funds, would come from the State and individuals joined by public advisors in corporate social responsibility, and resemble a digital transformation fund. Companies that establish a transition process for employees whose jobs are being replaced by AI could then receive subsidies (e.g. training with measures to encourage or ensure employee loyalty once training is completed).

Along the same lines, another potential solution was to create a fund to which both companies and workers contribute, which could lead to creating digital insurance (see "insurance mechanism"). In particular, this could potentially offset job insecurity by establishing a guaranteed minimum income.

Citizens also highlighted the need to review company structures to encourage including women (cross-sector consideridations), especially if the

future of work is in this field to mitigate risks of inequality. Citizens therefore suggested that funding be based on a points system that cultivates diversity (a type of diversity quotas for businesses supported by reinforcement policies, rather than sanctions).

Lastly, citizens liked the idea of developing a support program to create new business models for data processing businesses, such as co-ops. Their purpose would be to break the isolation of self-employed individuals, whose numbers will keep increasing.

Generally speaking, out of a political concern for sharing AI benefits and to ensure equitable distribution among social groups, territories and various vulnerabilities, participants recommend developing an AI development incentive policy that ties responsibility to business subsidies.

PARTICIPATIVE MECHANISMS

Participants suggested creating a multi-sector "permanent consultation space" within the government to respond to the division of powers (tied to the democracy principle) and address the challenges of how emerging sectors are structured.

Citizens also mentioned the importance of user participation in designing AI interface tools. They could be "design thinking" with different partners, and would allow them to review the work of the programmers, particularly to correct biases:

> "Allowing user input in machine learning through open AI (based on the Wikipedia model) to correct and review biases by and for society."

(Musée de la civilisation table, Québec City, April 6, 2018, AI as a compulsory step to employment scenario).

User feedback could be given to competent authorities (e.g. ethics committees, corporations) to adapt the system.

CODES OF ETHICS/CONDUCT

Two types of codes of ethics were suggested by citizens for the theme of workplace transformation: one for human resources advisors (CHRA) so recruitment efforts are carried out in unbiased fashion, the other for any profession using personal data for marketing purposes—such as advertisers—to ensure better protection of personal data.

The first, the CHRA code of ethics, would address the issue of "cultivating diversity through team building" and would be based on the results of a research program that studies recruitment biases and measures AI's impact on them (see "Research programs").

The second code of ethics would address the issue of protecting personal data. Participants call for "society to reflect on the use of personal data" in a context where they feel that the notions of "responsibility" and "common good" should be the subject of a democratic dialogue. This code of ethics would result from this societal and democratic thinking process and could be inspired by Europe's General Data Protection Regulation (GDPR).

> "Beyond individual consent (e.g. when visiting a website), we must reflect as a society on data use and issues of wealth redistribution."

DIGITAL TOOLS

A digital tool was suggested for the workplace sector: the creation of a personal digital file. This would consist of a unique portal to our digital data that obliges every business to declare the data it collects. This type of tool would have to be developed so that it operates transparently and intelligibly, particularly when using and storing personal data.

INSURANCE AND CONTRACTUAL MECHANISMS

To guide the digital transition and its impact on the workplace, citizens suggested creating digital AI insurance to allow each individual to become familiarized with AI and receive training on it. This insurance would be financed by a fund to which both workers and companies contribute (based on the same model as Quebec's parental insurance, adapted to the worker's reality). It could even facilitate access to training, and this training would be paid by companies (with an incentive measure, or even an employee loyalty program at the end of the training). Digital insurance could also help ensure a guaranteed minimum income to counter job insecurity for at-risk workers.

AI EVALUATION DEVICE

Impact studies were suggested to ensure that humans always come first in any AI system. These would be carried out by an independent organization funded by a data tax (based on the carbon tax model).

"When analyzing and creating any system, we must guarantee and maintain monitoring through an independent third party (if necessary), to put humans first. This organization would be funded by a data tax (like a carbon tax)."

(Musée de la civilisation table, Québec City, April 6, 2018, AI as a compulsory step to employment scenario)

RESEARCH PROGRAMS

Participants recommended developing multidisciplinary research programs that measure AI's impact on recruitment biases. In particular, this research program would inspire the creation of a code of ethics for HR advisors.

5.5.

HEALTHCARE

Table 8: Potential solutions or general guidelines for the healthcare sector

	Number of potential solutions formulated
Legal dispositions	11
Institutional stakeholders and other stakeholders	9
AI evaluation mechanisms	8
Training	6
Codes of ethics/conduct	4
Research programs	4
Professional frameworks and internal policies	2
Participative mechanisms	1
Digital tools	1
Total	46

LEGAL PROVISIONS

Several recommendations on rules and regulations were made at specific levels, particularly with regard to privacy, transparency, data collection and universal healthcare.

Many citizens felt that, although we can rely on existing laws and regulations when it comes a person's rights to control their personal data, we must also think of ways to redefine them to take technological innovations in AI into consideration. Protection of privacy was an important element in these discussions, and citizens expressed the need to guarantee the confidentiality of personal data.

"Laws should be introduced to guarantee private ownership of personal data (e.g. a law giving access to data collected from the people concerned)"

(INM table, Montréal, February 18, 2018, Digital twins scenario).

Certain tables also mentioned the need to implement laws and regulations that outlined clear and transparent objectives for collecting, using and accessing biological data (and any other personal health information). This information must be clear, understandable and readily available

to users. Participants highlighted the need to outline instructions for government organizations to provide intelligible, quality and relevant information when collecting personal health and biological data.

As to data collection, they also underlined the need to oversee sources used by the algorithm to ensure there are no biases against citizens. Participants also recommended introducing laws and regulations on the goals of the healthcare system to maintain a fair healthcare system, particularly in relation to the universal healthcare principle:

"Include all AI developments in healthcare in the law on access to universal healthcare, at the same level as alternative medicine"

(Mordecai-Richler Public Library table, Montréal, March 10, 2018, Vigilo scenario).

In the context of Canada, the suggestion was made to implement a law specifying if (and how) public healthcare coverage offered by the RAMQ could apply to technological innovations related to AI in healthcare.

Finally, participants also suggested that a regulation overseen by the College of Physicians be introduced to ensure that humans always come before AI.

"Robots must not be used without the supervision of a (human) institutional authority subject to a code of ethics."

(Mordecai-Richler Public Library table, Montréal, March 10, 2018, Vigilo scenario).

INSTITUTIONAL STAKEHOLDERS AND OTHER STAKEHOLDERS

Citizens recommended implementing many institutional committees and stakeholders in healthcare. These could be advisory committees whose mission would be to define the "values" that AI should consider when processing information. Citizens came up with the idea of creating an independent organization that could rule on privacy benefits and risks while also focusing on healthcare and ethical AI issues. Participants also felt committees needed to be established to review mistakes made by AI devices to improve algorithms. It could namely be requiring that the healthcare system periodically review the validity of its algorithms, and render public how they function and are evaluated with a "declaration of any modifications" clause". (Bibliothèque Père-Ambroise table, Montréal, March 3, 2018, Digital twins scenario).

Some participants thought someone should be designated as legally responsible so a human being is held accountable in the event of error. Likewise, a forum to appeal a decision made by an algorithm must be available. Establishing an independent ombudsman whose role would be to settle disputes between patients and doctors was also suggested.

Other citizens felt that appointing a life and well-being commissioner who "rules on healthcare objectives while defending citizens and the general population, and namely the right not to know" is crucial (Musée de la civilisation table, Québec City, April 6, 2018, Digital twins scenario). Creating a body to establish a humane and independent governance framework for AI development in healthcare was also suggested. Lastly, citizens recommended implementing a healthcare data anonymization centre managed by the government whose purpose would be to protect citizens from having their personal data misappropriated by private companies.

AI EVALUATION MECHANISMS

Citizens recommended establishing ethical AI certification in healthcare; namely developing a certification (or label) for algorithms and robots from research project databases (participative study on what influences AI development) to determine the criteria and various levels of this certification. The criteria should include transparency, security and relevance of the tool. For example, these certifications would be designed to standardize access to the algorithmic decision-making process, or validate the tools of healthcare robots. These certifications should be issued by the government

or independent, multiparty organizations to protect public interest and patient well-being, and mainly target private companies developing AI healthcare.

"Upfront certification for healthcare robots and their tool kits (particularly to protect public interests)"

(Mordecai-Richler Publid Library table, Montréal, March 10, 2018, Vigilo scenario).

TRAINING

Participants recognized the need to establish education and awareness measures for all stakeholders involved in AI development for the healthcare sector including healthcare professionals and the public. Professional training, which could be in the form of ongoing training (e.g. based on creating a best practices guide) should particularly focus on the doctor-patient-AI relationship, with case studies and updated statistics. The purpose of this training would not only to make an optimized and informed use of algorithms, but provide adequate and accurate communication of information to patients to avoid misinterpretation.

As for public training, participants recommended that awareness begin from day one of the younger generation's education (in school) to cultivate critical thinking about AI technologies. Citizens proposed the idea of an intellectual self-defence class to develop critical awareness and educate users about new practices through outreach.

"In primary school, start raising awareness among younger generations and cultivating critical thinking. Ensure information shared with the public is accurate and determine what information deserves to be shared with citizens/patients."

(Musée de la civilisation table, Québec City, April 6, 2018, Digital twins scenario).

CODES OF ETHICS/CONDUCT

Citizens also recommended adopting codes of ethics, whether for any company creating AI for healthcare, or more globally for Canadian users and healthcare professionals. These codes must contain standards as to the safety, transparency and responsibility of doctors or developers. These codes should help ensure that every citizen is accompanied by a doctor for any medical decision. Some citizens mentioned that the definition of human responsibility toward AI needed to be added to existing codes of ethics. For example, it was suggested that a Hippocratic oath 2.0 be implemented. This would ensure that people receive personalized care and monitoring by including healthcare professionals in all healthcare recommendations. This could involve implementing "virtual guardrails" to prevent the algorithm from going off-track and skewing the diagnostic.

"The doctor's responsibility and code of ethics should always prevail over AI. AI is just a tool to help."

RESEARCH PROGRAMS

Citizens recommended establishing, funding and fostering various multidisciplinary research programs on AI in healthcare. Participants all agreed that AI research should be at the forefront, but so should other disciplines that study the effects of AI on society, such as social sciences, philosophy or bioethics. These studies should, for example, help identify shared responsibilities among the various stakeholders, measure the impact of AI on their autonomy or launch training and education programs for both practitioners and citizens.

"Develop research programs to evaluate the degree to which an individual's socioeconomic status has an impact on their health and eventual AI diagnosis"

(INM table, Montréal, February 18, 2018, Digital twins scenario).

PROFESSIONAL FRAMEWORKS AND INTERNAL POLICIES

In response to the risk of attacks on privacy, citizens recommended that the healthcare system be responsible for documenting and informing patients when their data is accessed by third parties ("who" and "when").

Citizens also recommended a procedure to follow for a diagnosis (in the same vein as a combined human-machine diagnosis). This procedure would encourage doctors to make a diagnosis before the algorithm, which would help safeguard the doctor's expertise and independence, and ensure the algorithm remains a complementary tool to inform the doctor and assist them in decision-making. This algorithm would not only strictly consider a patient's medical data (e.g. biological indicators), but other kinds of data (e.g. lifestyle, eating habits).

PARTICIPATORY MECHANISMS

Citizens highlighted the need to hold a debate and public consultation on data safety before introducing one or many bills. These debates should include the public, experts and other stakeholders who are already involved (e.g. ethicists).

"We have to go beyond the context of an ordinary citizen on their computer to dealing with a privacy policy."

DIGITAL TOOLS

The creation of an electronic consent form adapted to the digital reality was suggested. It should be user-friendly, digital and interactive, and a contact person should always be available to consult.

5.6.

SMART CITIES AND CONNECTED OBJECTS

Table 9: Potential solutions or general guidelines for the smart city and connected objects sector

	Number of potential solutions formulated
Legal provisions	14
Institutional stakeholders and other stakeholders	10
Training	10
AI evaluation mechanisms	5
Participative mechanisms	5
Digital tools	3
Codes of ethics/conduct	2
Incentives	1
Research programs	1
Total	51

LEGAL PROVISIONS

Participants at tables discussing the smart cities and connected objects theme suggested implementing a number of legal provisions. The goal of these potential solutions would be to protect personal data and user consent and guarantee the loyalty of technology. For example, citizens suggested a regulation authorizing disconnection at any time as a means to control connected objects. Also, in response to various risks (including invasion of privacy), participants invited people to consider including a legal provision on the loyalty of connected objects, which would guarantee that the measures taken and the recommendations made are in the interest of the consumer, not the company:

> "Law defining the notion of loyalty and other ethical considerations (discrimination)"

(SAT table, Montréal, March 13, 2018, Smart toy scenario).

Citizens recommended legally determining an age for "digital maturity" for use of technology by minors:

"We have to think about an age for digital reasoning. About digital maturity."

This measure echoes the suggestion to cultivate "digital citizenship", which would help empower citizens to deal with changes dictated by new technologies. This would help define responsibilities and educate citizens on their rights and responsibilities regarding AI accessibility, in particular.

Citizens also came up with the idea of introducing a moratorium. It could last one or two years and would help provide a legal framework for the use of artificial intelligence in public transportation:

"Prior to implementation, we must set some parameters. We need to impose a moratorium until we have responsible technology."

For equity issues, participants suggested establishing a mobility social assistance program that would help remove barriers to AI access for certain at-risk categories of people. Likewise, citizens recommended establishing a right to mobility that ensured everyone access to transportation. Reforming transportation laws, traffic regulations and road safety was therefore suggested. Citizens also felt that urban planning laws needed to be reviewed; for example, by introducing regulations that promoted mixed development and took population diversity into account.

Establishing regulations to help secure personal data and information sharing was also recommended. These regulations would help protect anonymity and data ownership, ensure the proctection of privacy or prohibit data capture outside of planned hours. These laws should also provide for greater transparency in the handling of personal data by the private sector.

"Broaden the scope of the law on consent to guarantee that individuals maintain ownership of their own data."

(Musée de la civilisation table, Québec City, April 6, 2018, Connected refrigerator scenario).

Citizens believe that these laws need to be integrated into the Constitution of Canada. To protect users' transportation parameter choices, citizens suggested introducing federal laws while maintaining regulations that could be adapted to the local level.

INSTITUTIONAL STAKEHOLDERS AND OTHER STAKEHOLDERS

Table participants discussing the theme of smart city and connected objects came up with many ideas for creating institutional stakeholders, whether independent societies or advisory committees. The democratic ideal of committees or assemblies that allowed for citizen participation was brought up many times.

For the control of connected objects, two models were suggested, including a mechanism where private stakeholders would be forced to self-regulate:

> Based on the model of the Régie du logement du Québec, a **Régie des objets connectés** (Connected Objects Board) would help set prices for connected objects (such as refrigerators) and would provide social assistance to help people buy them. It would also issue ownership certificates when purchasing a connected object, to confirm that the data generated by this object belongs to the user. This person could then choose whether they consent to the data being communicated to the marketing and insurance company, without any risk of penalty.

> An independent data management authority would allow citizens to file a class action suit in the event of harmful use. It could also manage a digital platform where users could speak freely and publicly about the advantages

and disadvantages of AI devices and thereby have an impact on the brand image of private stakeholders marketing these devices. These private stakeholders would then be forced to self-regulate through user pressure on their brand (Musée de la civilisation table, Québec City, April 6, 2018, Connected refrigerator scenario).

To respond to the issue of equity and thereby ensure an equitable sharing of AI, an advocate could be reached at "1–800 discrimination of connected objects" (INM Table, Montréal, February 18, 2018, Connected refrigerator scenario). It could then be a part of a "multiparty committee that democratically oversees incidents, injustices and other issues" (Mordecai-Richler Public Library table, Montréal, March 10, 2018, Self-driving car scenario). Furthermore, an independent auditor could be mandated to lead an accounting audit to ensure equitable sharing of AI benefits (INM Table, Montréal, February 18, 2018, Connected refrigerator scenario).

For self-driving car regulations, the creation of the SAIAQ (Société de l'Assurance de l'Intelligence artificielle du Québec) would introduce changes to road safety laws to include autonomous driving. It would also provide auto insurance 2.0 that would offer new kinds of contracts for this type of driving (Bibliothèque du Boisé table, Montréal, March 17, Self-driving car scenario).

To organize smart city networks efficiently and optimize the urban system managed by AI, participants suggested a hybrid organization: the MAIUO (Mobility, artificial intelligence and urban optimization) funded by the Quebec government (SAT table, Montréal, March 13, 2018, Self-driving car scenario). This centre's mission would be to manage and optimize the engineering for AI and pool knowledge to help draft laws and regulations following pilot projects.

Participants also considered training different groups of people, such as a Minister of Technological Development that would advise the Minister of Smart Territories, which in turn would provide a framework for urban changes related to AI and sustainable cities; or even a commission to defend the right to mobility for self-driving vehicles to guarantee protection of the right to mobility (see Legal provisions).

TRAINING

Participants recommended implementing training for citizens on new technologies and smart cities, so they could gain a better understanding of how AI operates and the new standards that come with it. This education could be in the form of outreach, ongoing training or awareness campaigns. It could, for example, focus on AI operations and use, or civic life and the digital city.

Participants recommended collective vigilance training for responsible AI use. This training would democratize AI information to educate individuals on its rules of use, cultivate informed choices and allow them to take part in the decision-making process.

"Data literacy courses offered at different levels of education to provide citizens with the tools and reflexes to make informed choices"

(SAT table, Montréal, March 13, 2018, Smart toy scenario)

AI education in the city sector must occur at every level and in different locations (e.g. library, co-op, fab lab, school or non-profit organization). It could be a hands-on course in schools to teach students how to manage different connected objects, or digital literacy education programs.

AI EVALUATION MECHANISMS

Citizens recognized the need to implement mechanisms to evaluate the costs, side effects and impacts of AI-specific policies. They considered establishing standards (e.g. ethical labels) to protect the consumer, put human beings first in decision-making and foster inclusion. For example, citizens suggested establishing an ISO-like certification that would recognize companies that offer digital services with added value for citizens. This standard would guarantee that users' control their choice of services to prevent services from becoming intrusive.

Creating a certification that ensured collaboration between humans and machines was also suggested. It would guarantee user safety, security, operability, transparency, loyalty and/or trust:

"Certification that measures and guarantees the level of loyalty and other ethical considerations of my connected object"

(SAT table, Montréal, March 13, 2018, Smart toy scenario).

PARTICIPATORY MECHANISMS

Citizens recommended implementing public assemblies such as hybrid democratic forums so that citizens could evaluate projects and user needs, and determine how public spaces are planned according to people's needs and society's values. Citizens also suggested implementing a class action system for abusive use of AI, which would be dynamic, flexible and able to adapt to technological progress.

Other suggestions involving active citizen participation were presented, such as introducing surveys and participative planning (evaluating urban planning during the transition period), systems, even an open-source code of ethics (to find solutions to community issues and improve community well-being). Citizens highlighted the need to review jurisdiction between the province, municipalities and districts.

DIGITAL TOOLS

Participants suggested integrating a type of development into the design of connected objects that would allow users to easily understand and visualize the data generated by objects (who/when/where they are sending it and why), to ensure that they could easily customize their settings. The idea would be to ensure a multidisciplinary design of connected objects that integrates the emotional and psychological aspects of an individual's relationship with food or other elements into the design process (see Connected refrigerator scenario) or recommend travel options based on personal criteria (see Self-driving car scenario).

CODES OF ETHICS/CONDUCT

Citizens also recommended introducing a code of ethics for computer engineers and AI designers, which could be implemented and monitored by an independent organization. It would rule on the need for transparency and traceability, inclusion and factor in risks to protect the public. This code would be a responsibility permit to protect the common good.

INCENTIVES

Citizens recognized the need to establish incentives to encourage companies to reveal their sources and biases, the algorithms they use and ensure the transparency of recommendations and actions of connected objects (e.g. through tax breaks or calls for tender). These incentives (whether individual or collective) could also encourage the use of other means of transportation (see Self-driving car scenario). For example, these incentives could be a mobility points system for individuals who use shared transport, especially that which runs on green energy or has low greenhouse gas emissions.

RESEARCH PROGRAMS

Participants highlighted the need to conduct studies to understand the implications of AI use and guarantee the harmonious development of society at various levels as well as reflect on preserving human heritage.

"Conduct studies to understand the implications of AI use and guarantee the harmonious development of society (psychology, culture, social issues, equality, education"

(SAT table, Montréal, March 13, 2018, Smart toy scenario).

They also suggested establishing pilot projects that promoted public transit in the city and took social equity issues into consideration, while helping eliminate design barriers.

INSURANCE MECHANISMS

Although these recommendations did not make their way to the posters, participants suggested implementing digital insurance to ensure integrity and protect ownership of personal data, whether for self-driving cars or connected objects. Moreover, creating new types of contracts for automobile use was suggested to ensure proper AI management for individual mobility.

6. CONCLUSION

A number issues and potential solutions were identified as a result of this deliberation workshop which brought together hundreds of citizens, whether enthusiasts, users or experts. The goal was to listen to what citizens had to say about responsible AI development, and discussions were organized around scenarios that showcased the many risks and various ethical issues that had been identified ahead of time, echoing the Declaration's principles. These observations should help overcome skepticism of AI development which may emerge from these results, without necessarily ignoring it. The results give us a certain idea of the social acceptability of AI and its development.

The wide range of suggestions implies that we deepen the analysis to make recommendations for public policies. All results presented raised a number of issues, which must be analyzed further in order to formulate these recommendations. Focusing on these issues appeared crucial to issue a statement on a responsible framework for AI development. They are discussed in the following priority projects of this report:

1. Addressing the challenges of AIS **governance**
2. Developing **digital literacy** for all citizens
3. Ensuring **diversity** in AIS development
4. Promoting strong sustainability AIS development

AI development therefore raises many societal issues. Although these challenges are not all necessarily specific to AI, the transformations caused by its development in various social spheres call on us to question ourselves as citizens and on the society we wish to build. At the heart of this tension between hope and fear, it is the relationship between humans and technology that needs to be highlighted. If one request seems to be unanimous, it is ensuring that humans remain front and centre in a world that is increasingly becoming artificially intelligent.

< >
Montréal Declaration
Responsible AI_
</ >

PART 4

FALL 2018 CO-CONSTRUCTION: KEY ACTIVITIES

TABLE OF CONTENTS

1. INTRODUCTION 152

2. CO-CONSTRUCTION DAY OUTSIDE QUEBEC (PARIS, FRANCE) 153

2.1 Addressing democracy issues raised by fake news 156

2.2 Discussing environment-related issues 160

2.3 Addressing issues around digital transformations in the workplace 164

3. DISCUSSION OF THE CULTURE THEME WITH MEMBERS OF THE COALITION FOR THE DIVERSITY OF CULTURAL EXPRESSIONS (CDCE) 171

3.1 Three themes proposed by the Declaration team to facilitate debate on AI development issues in the field of culture 171

3.2 Promoting cultural diversity in the AI era 172

3.4 The CDCE's ethical principles 174

3.5 Selected recommendations 176

4. BRIDGING THE GAP BETWEEN CITIZENS AND A NEW GENERATION OF RESEARCHERS: POLICY BRIEF SIMULATION 177

4.1 Description of the activity 177

4.2 Problems identified on the basis of citizen concerns 179

Problem 1. Public Security and System Integrity 179

Problem 2. AI, the Media and the Manipulation of Information 180

Problem 3. Public, Private or Participative Governance: Digital Commons 181

4.3 Recommendations from the new generation of researchers 182

5. CONCLUSION 184

Appendix 1 – The Paris Scenarios (in French only) **185**

Démocratie 185

Environnement 186

Monde du travail 187

Appendix 2 – Student policy briefs 188

TABLES AND FIGURES

Charts 1, 2, 3 and 4: Profile of Participants in the Co-Construction Day in Paris 153

Diagram 1: Issues Raises and Potential Solutions Proposed at the Co-Construction Day in Paris 155

Table 1: Democracy, First Deliberation Period: Identification of Ethical Issues in 2022 158

Table 2: Democracy, Second Deliberation Period: AI Management proposals for 2018-2020 158

Table 3: Environment. First Deliberation Period: Formulation of Ethical Issues in 2025 162

Table 4: Environment, Second Deliberation Period: AI Management Proposals for 2018-2020 163

Table 5: Priority Issues 167

Table 6: Proposals Retained 169

Chart 5: Profile of Participating Students, Based on Area of Study (according to the three FRQ funding areas) 178

1. INTRODUCTION

The Montréal Declaration's main co-construction activities were carried out from November 3, 2017 to April 31, 2018. The Declaration team nevertheless continued to work on the project in the fall of 2018, organizing three key activities to mobilize the knowledge of more actors on various important issues in responsible artificial intelligence (AI). A day of co-construction was organized in Paris using the winter 2018 co-construction model. To mobilize the knowledge of stakeholders in the cultural sector, a focus group was also held on issues related to the advent of AI in fields related to art and culture. Lastly, as a bridge between the public consultations held last winter and ongoing research, an activity was carried out with graduate students, simulating the drafting of policy briefs, in partnership with the *Comité intersectoriel étudiant* (CIÉ) of the *Fonds de recherche du Québec* (FRQ).

These activities supported further analysis and the drafting of recommendations on public policies (see Part 6 of the report *Priority projects and their recommendations for responsible AI development*). The following sections provide a recap of the issues identified as well as the main potential solutions developed by the participants in these activities. Some draw on the mechanisms proposed during last winter's co-construction, and support the need for their implementation, while others add new elements to the discussion.

WRITTEN BY

NATHALIE VOARINO, Scientific Coordinator, PhD Candidate in Bioethics, UdeM

CHRISTOPHE ABRASSART, Associate Professor in the School of Design at the Faculty of Planning, UdeM

CAMILLE VÉZY, PhD Candidate in Communication Studies, UdeM

IN COLLABORATION WITH

LOUBNA MEKKI BERRADA, Doctoral student in Neuropsychology, UdeM

VINCENT MAI, Doctoral student in Robotics, UdeM

2. CO-CONSTRUCTION DAY OUTSIDE QUEBEC (PARIS, FRANCE)

This section presents results from a co-construction day held in Paris on October 9, 2018, organized in partnership with the Canadian Embassy in Paris, the Canadian Cultural Centre and the House of Canadian Students. At this event, 26 persons of varied backgrounds were mobilized to examine issues related to responsible AI. The participants were assigned to one of three co-construction tables, with each table addressing a key theme in AI development: Democracy, the Environment and the World of Work.

Chart 1: Profile of Participants in the Co-Construction Day in Paris

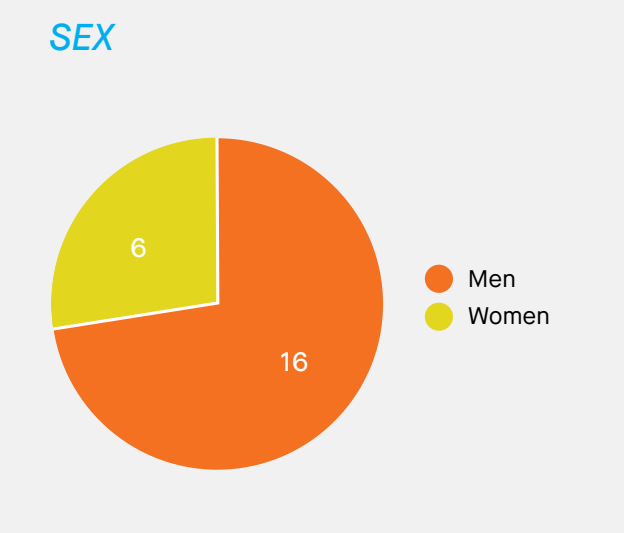


Chart 2: Profile of Participants in the Co-Construction Day in Paris

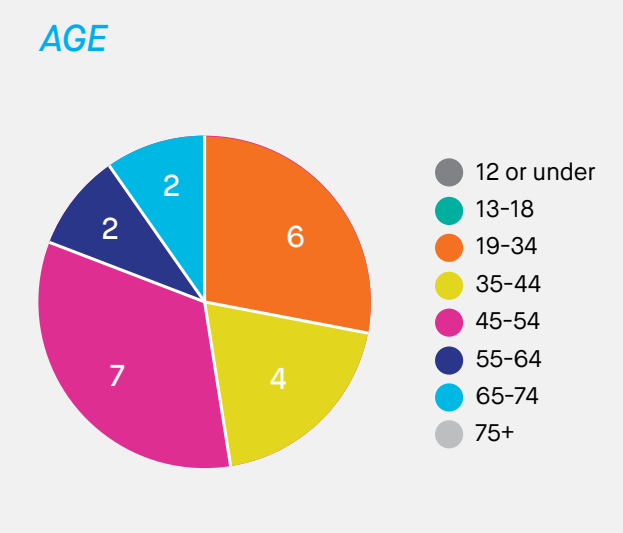


Chart 3: Profile of Participants in the Co-Construction Day in Paris

EDUCATION | (out of 22 respondents) 2 respondents gave no answer

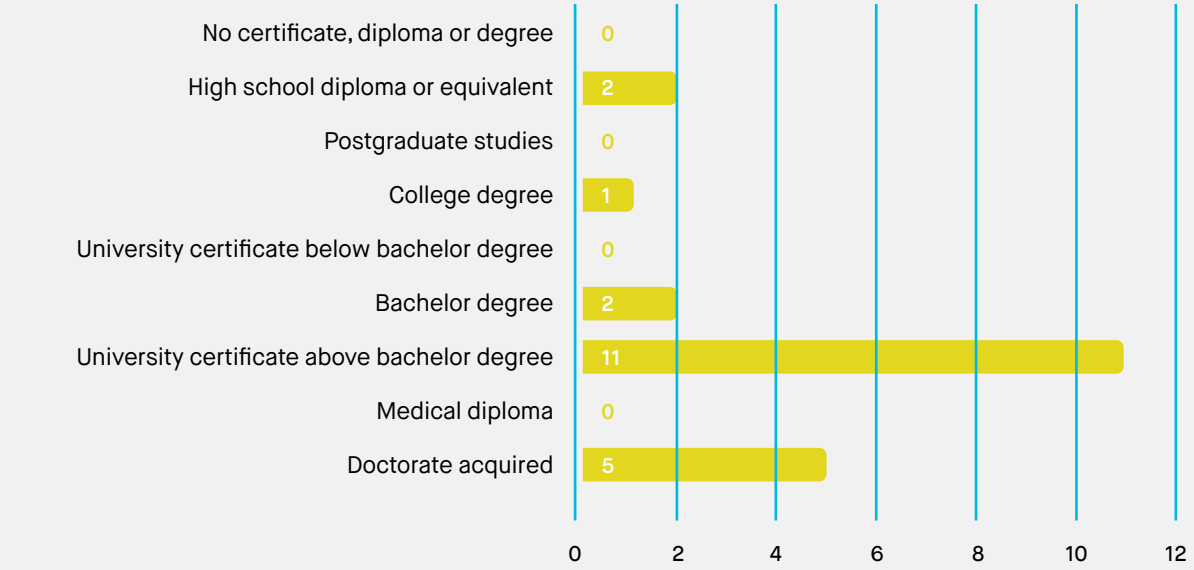
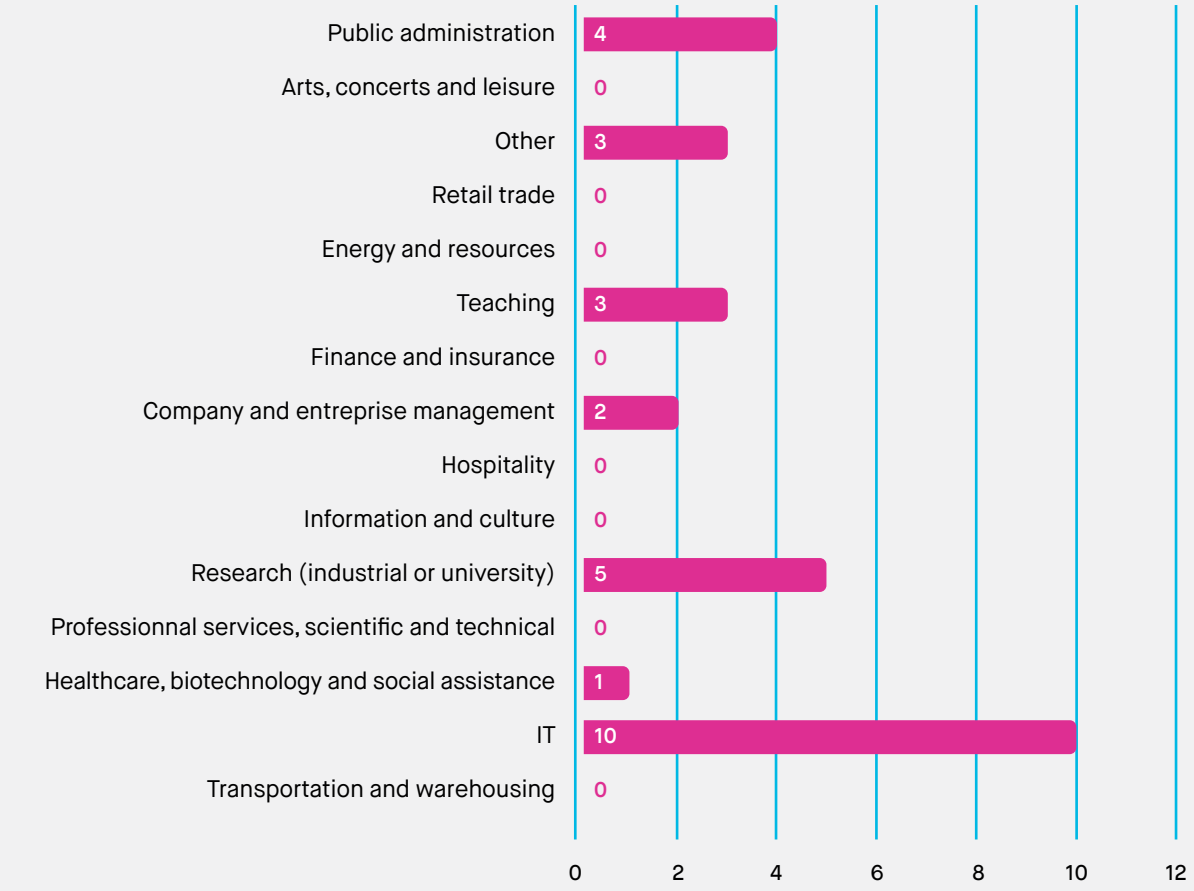


Chart 3: Profile of Participants in the Co-Construction Day in Paris

FIELDS OF ACTIVITY | (out of 22 respondents)

7 respondents indicated more than one sector of activity; 2 respondents gave no answer



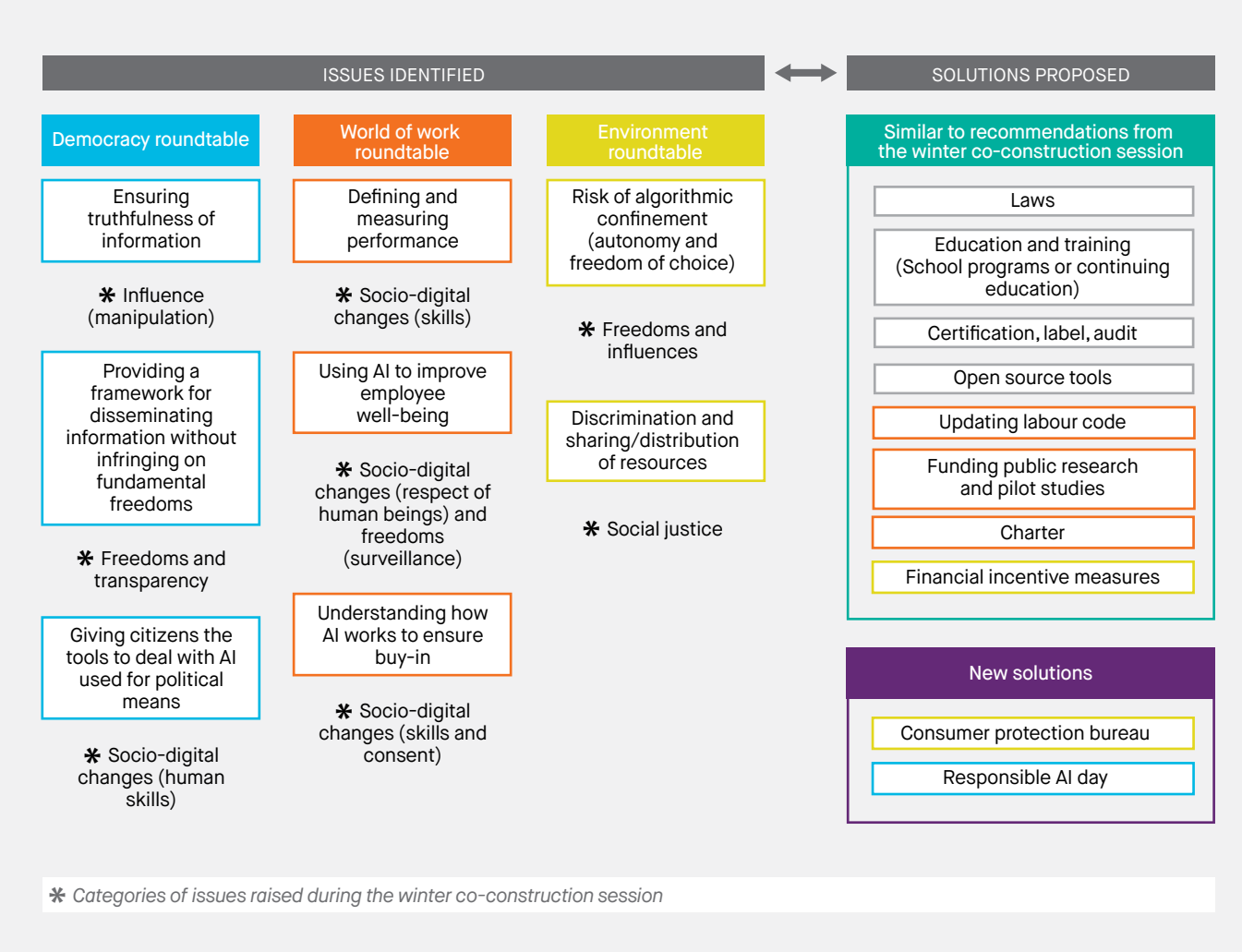
The discussions were organized around three distinct deliberation periods (identifying the issues, preparing recommendations and writing a front-page newspaper article), following the co-construction model developed for last winter's activities. Three trigger scenarios were used (see Appendix 1, in French only), involving the use of AI:

1. to encourage ecological behaviour,
2. in human resource management in a business, and
3. to create fake news during an election campaign.

These three scenarios allowed us to explore issues related to the environment, the world of work and democracy from a new point of view, since we had never used this format before.

The following sections relate the directions taken by the discussions at each of the three discussion tables[1]. They highlight the appearance of issues that had already been identified at last winter's co-construction activities, but in the specific given contexts. In addition, various potential solutions or mechanisms for managing responsible AI were proposed. Some of them are similar to those formulated at last winter's activities (e.g. Laws and Training), while others are new (e.g. Responsible AI Day).

Diagram 1: Issues Raised and Potential Solutions Proposed at the Co-Construction Day in Paris



[1] The citations were taken from post-its written by the participants.

2.1.

ADDRESSING DEMOCRACY ISSUES RAISED BY FAKE NEWS

Summary of the initial scenario

Fake News During an Election Campaign

Two weeks before a presidential election, an emergency meeting is called by the Information Integrity Agency (IIA), which was implemented under the *Act to Combat the Manipulation of Information*. A video has emerged showing the outgoing president making compromising remarks on immigrant workers, and it has gone viral. The president's spokesperson says that the video is fake, created by a foreign agency trying to interfere with the elections, using GAN algorithms (Generative Adversarial Network). Even though dissemination of the video is prohibited, it continues to circulate on various foreign websites. With just one month to go before the first round of the presidential elections, the IIA must come up with a plan on how to contain the devastating impact of this misinformation and restore conditions for a healthy campaign.

The objective of this scenario was to stimulate a discussion of ethical issues related to information manipulation, which can harm democracies when it spreads virally. This is particularly true when artificial intelligence techniques are used to imitate people and change content while maintaining a very high level of realism, making the detection of what is fake very difficult.

The deliberations presented here are the result of a full day of discussions among seven researchers, experts and students working in the fields of ethics, organizational development, machine learning, the social web and political science. Taking this scenario in 2022 as a point of departure, the discussions led to drafting a headline and lead for a front-page article in the responsible AI newspaper, dated October 9, 2022: "First Responsible AI Citizen Day."

First Discussion Period: FORMULATION OF THE ETHICAL ISSUES IN 2022

DEMOCRACY

The participants believe that AI itself is not the cause of the attacks on democracy. These problems already existed; however, they are altered or aggravated by the opportunities afforded by AI. Consequently, we must prepare for this new reality, and make the necessary adjustments. It was therefore suggested that "maintaining a healthy democracy" is a particularly important challenge, in a context where "choices based on wrong information" cause problems. A legal framework and training in critical thinking were discussed as two ways to defend against the impacts of manipulated information and maintain a healthy democracy.

LEGAL FRAMEWORK

The participants wondered how practices related to the manipulation of information could be controlled:

> "Can the law effectively control all AI practices (manipulation of information)?"
> - A participant

In this context, one question came up several times: "how can justice be brought to bear" when one's reputation has been sullied by fake news? One of the impediments is that it is difficult to control the manipulation of information without limiting freedom of expression and other fundamental freedoms.

KNOWLEDGE AND CRITICAL THINKING

"The importance of the interdependency of democracy and education (in particular, education in critical thinking)" was mentioned in a call to develop critical thinking for *"informed participation in public life."* One participant said that instead of mounting a defence against propaganda, it would be better to develop critical thinking, which allows individuals to defend themselves against the impacts of propaganda (indoctrination, changes in choices and behaviour, extreme polarization, etc.). Preventing all propaganda, i.e. all actions taken to influence opinion, could lead to censorship or curtail freedom of expression. Rather we should focus on educating citizens to develop "critical thinking" and "media and statistical literacy."

The issue of **democratizing access to information and knowledge** was then raised. Several conditions must be met: net neutrality must be ensured, in order to guarantee free access to all information, and everyone must have access to technological tools, so that they can obtain information and express themselves. Everyone, without discrimination, must therefore be able to learn about how AI works and its related issues.

AUTHENTICATION AND AI'S ABILITY TO IDENTIFY WHAT IS FAKE

By learning certain AI techniques, some may become able, for example, to develop tools for detecting and correcting fake news. However, this depends on the potential and the need to use AI to sift out the fake from the real in a context where, for example, it is impossible to distinguish a real video from one created by AI with the naked eye. The participants therefore examined "mechanisms for authenticating the news/information" and "technology's ability to understand sarcasm." One example given of a potential authentication solution was the "traceability of sources on dissemination tools (e.g. WhatsApp)." It was nevertheless mentioned that an AI that can automatically censor publications it flags as fake, malicious or unreliable would curtail freedom of expression.

RESPONSIBILITY

Given the impression that "control over information" is lost when fake news spreads very quickly or even goes viral, participants discussed the issue of responsibility in several different stages: creating, disseminating, sharing and reading information. They wondered *when* the truthfulness of the news should be evaluated (before it is created, before it is published, before it is shared, or when it is read?), *by which entity* (an international organization, the state, journalists, readers, the dissemination platform used to publish the news and/or sharing on platforms such as WhatsApp and Facebook?), and *how* (developing the reader's critical thinking, creating a label indicating the information's level of truthfulness and the source's reliability?).

The creation and dissemination of fake news are also seen as combining the roles of several of the actors needed to take on the responsibilities of creating information, disseminating it, determining its credibility, and confirming that its sources are truthful and reliable; i.e. journalists, readers, the entity that should verify the truthfulness of the news, the information dissemination platform and any individual independently creating or disseminating information.

Participants discussed the "ethical conduct of the media (capturing and disseminating information)" and "journalism's credibility." A new role for journalists and the media could be to judge the information published and shared, checking it and declaring whether it comes from a reliable source.

Table 1: Democracy, First Deliberation Period: Identification of Ethical Issues in 2022

Ethical issues in 2022	1	2	3
Description	Ensure that the information is truthful	Manage the dissemination of information without infringing on fundamental freedoms	Equip citizens to deal with the political uses of AI
Related principles	Democracy, responsibility	Democracy, autonomy	Democracy, knowledge, autonomy

Following this discussion, the participants formulated three priorities for which potential frameworks need to be proposed to keep democracy healthy, despite the impacts of propaganda in the form of the automated creation and dissemination of misinformation. Those three priorities are:

1. To ensure that the information is truthful and reliable, in particular to preserve the health of democratic discussion.
2. To control the dissemination of information without infringing on fundamental freedoms, including freedom of expression, in particular through the development of journalistic and technological standards on the dissemination of information.
3. To equip citizens to understand the political uses of AI so that they can learn about them and freely develop their own opinions.

Second Discussion Period: PROPOSALS FOR OVERSIGHT OF AI

In response to these issues, the team's discussions led to a series of five proposals on AI oversight:

Table 2: Democracy, Second Deliberation Period: AI Management proposals for 2018-2020

Management proposals	1	2	3	4	5
Description	Creation of an authority that would certify journalistic standards with a label (red, yellow or green label)	Creation of open-source tools that can distinguish between true and false (e.g. mobile phone app)	A bill to guarantee net neutrality	Implementation of school and continuing education programs (such as MOOC) for developing critical thinking skills	Implementation of a Responsible AI Citizen Day
Related issue	Ensuring the truthfulness of information	Control the dissemination of information without infringing on fundamental freedoms		Equip citizens to understand the political uses of AI	

In line with the previous ideas, the discussions that led to these proposals identified the technical, political and educational dimensions of the problem of manipulated information. In response, the participants proposed a series of measures intended to structure fact-checking of the news (1, 2) and educate citizens for free and informed participation in democratic life (3, 4, 5). This will require standardizing information creation and dissemination practices (certification) and fake news detection support (tools), equal information access for all through a neutral Internet network (legislation), education for all—at every stage of life—to develop critical thinking (school and continuing education programs), and raising awareness about the political uses of AI (responsible AI day).

Brainstorming on the responsibilities of journalists resulted in a proposal for a new certification authority that would establish journalistic standards through a system that includes an information reliability indicator. This would serve as a way to establish the reliability of sources and the credibility of information. The participants felt that such a certification organization should be independent of government.

Instead of a legal framework, the participants preferred an approach based on professional standards and the establishment of guidelines on how to create, fact-check and disseminate information. One participant pointed out that a minimalist approach would be more reasonable for the design of a system that indicates reliability of information, to avoid a situation where the system has excessive influence over electoral processes and other competitive situations in which the news plays a crucial role. For example, this might happen if the reliability indicator system could give one party an advantage, or be used in a strategy to manipulate or lobby against an adversary. At the same time, a minimalist approach would be better placed to avoid excessive limits that could infringe on fundamental freedoms.

In opting for this preventive solution, the participants did not propose any solutions on how to respond to the dissemination of fake news. They nevertheless mentioned that when this occurs, there would be a need to intervene in its dissemination as quickly as possible, and refute the fake news with the fact-checked version.

The fact-checking journalistic practices adopted by this certification authority could employ an open-source tool whose technology would find fake elements even when they are undetectable to the naked eye, in particular when they have been created using AI techniques. The practices should also check to ensure that certain accurate information is not identified as fake news.

The participants also pointed out the large share of the responsibility that needs to be shouldered by information-sharing platforms such as social media (Facebook, Twitter, YouTube, Snapchat, etc.) and instant messaging platforms (e.g. WhatsApp). They wondered about the need to oblige these platforms to install a tool that uses AI to identify fake news. The participants did not appear to trust private businesses to install a fair system for identifying and blocking fake news, given the fact that, for now at least, this type of platform does not flag fake news videos that are influencing the opinions of various audiences. However, it would be a positive step if these platforms had a system that would indicate the level of reliability of a news item, since they absolutely must be held accountable for the role they play in disseminating manipulated information affecting democracies.

Training citizens and raising their awareness are subject to a second series of measures that target the development of critical thinking among citizens, in particular through media literacy that will equip them to freely browse the information universe in an informed manner. This could be achieved through school curricula and public spaces available to all, through public universities in libraries or cafés, for example, or even through creative awareness-raising campaigns that reach people in their day-to-day lives to keep them vigilant about information manipulation practices.

Third Discussion Period: WRITING A HEADLINE AND LEAD FOR A FRONT-PAGE NEWSPAPER ARTICLE FOR 2020

These proposals were then made into a narrative in a headline and lead of the responsible AI newspaper for October 9, 2022, as follows:

First Responsible AI Citizen Day

Several events were held simultaneously in Canada and France as part of the first Responsible AI Citizen Day, particularly to inform citizens about the new opportunities offered by AI and the importance of being able to think critically and to equip them accordingly. To this end, we are very proud to inform you that our paper has been accredited by the Order of Responsible AI Newspapers.

By promoting a "Responsible AI Citizen Day," the participants underscored the importance of raising awareness among all citizens of the need for them to appropriate AI issues to be able to participate in democratic life. The special mention about the paper being accredited as a responsible AI newspaper also reinforces the role played by the media and journalists as important actors in healthy democracies.

2.2. DISCUSSING ENVIRONMENT-RELATED ISSUES

Summary of the 2025 scenario given as a point of departure

The mercury keeps rising, with record-breaking temperatures recorded all over the world. In response to the climate change crisis, various cities are talking about introducing **EcoFit**, a highly incentive-based individual carbon permit system that is connected to the citizen's bank account and various online shopping apps. In these cities, the prices of goods and services are posted in euros and carbon, and each citizen must aim for a total personal consumption representing a maximum of 4 tonnes of carbon emissions per year. This rating gives them access to a series of environmentally responsible transportation, education, professional development and cultural services. In 2025, Ive and Charles have managed to gradually adjust their consumption to meet this target, and even get below it. And since they have spent less, they have saved more than expected. So they are considering a Christmas vacation in Cuba, and have begun visiting travel websites. But then they receive a message on their phone: "Beware the rebound effect: spending your savings on a flight will negate all your hard work. Think about planning a trip closer to home!"

The objective of this scenario was to encourage a discussion of what could be achieved on ethical issues through predictive management (using AIS) of environmental rebound effects on the consumer and equipment markets. Rebound effects can be explained as follows. As equipment becomes increasingly energy efficient and the environmental footprint of consumer goods becomes smaller

through better eco-design, we tend to consume proportionally more equipment, goods and services. This means that our gains are lost rather than locked in. For example, we buy larger screens, fill our homes with more equipment, travel farther in our cars, travel by air, etc. The result is an increase in GHG emissions and even greater pressure on resources and biodiversity. Given these rebound effects, economic development must be considered alongside its material reality and ecological footprint.

The algorithmic apparatus imagined here is part of an "AI for Earth" approach to using AIS. This exploratory scenario also includes predictive and tailored management of rebound effects, through supervised learning based on consumption histories (e.g. bank transaction histories) that are associated with nudges.

The deliberations presented here are the result of a three-hour round-table discussion involving eight citizens with an interest in new technologies and in environmental and sustainable development issues. The discussions based on this scenario for 2025 led to the formulation of an initiative presented in a headline and lead of the responsible AI newspaper dated October 9, 2020: *"Major Achievement for ConsoM'AI: 1 Million Subscribers in a Week. General Regulation on Opening Data for the Environment."*

How did the group's deliberations lead to this original proposal? What were the pivotal moments in their discussions? How did the ideas develop in each stage? The following sections present some key moments from this group's deliberations along with our comments.

FIRST DELIBERATION PERIOD: FORMULATION OF ETHICAL ISSUES IN 2025

During the event, participants jotted down their questions about the various principles underlying the Montréal Declaration on a series of Post-its:

PRIVACY PRINCIPLE

"Can we reconstruct this family's entire consumption history?" "Will the database be managed reliably enough to protect personal data and win users' trust?" "Could there be a right to erase?"

AUTONOMY AND FREEDOM OF CHOICE PRINCIPLE

Does this AI application lead to a new "prescriptive power?" "Does it maintain autonomy in decision-making and free will?" "How can one think critically about these personalized recommendations?" "Does this represent a machine exercising control over day-to-day life?" "Is there a risk of algorithmic confinement, of algorithmic bubbles?" "How can such a system take into account the singular context of a purchase decision (e.g. an emergency situation)?"

RESPONSIBILITY PRINCIPLE

This measure needs to help "strengthen environmental responsibility in day-to-day life," and "to express personal ethics as an actor in consumption." But, when relying on AI tools, "is there a risk of externalizing personal responsibility?"

JUSTICE AND EQUITY PRINCIPLE

By asking businesses to assess the carbon footprint of their products and services before they enter the market, does this application "ensure free competition?" "Is there a risk that it will expand the market power of large corporations and discriminate against SMEs by creating a barrier to entry, due to the cost of these environmental assessments?"

"How will fair trade, which has other ethical dimensions, be evaluated?" "Will some producers be favoured over others?" "The rich will be able to consume more and buy carbon quotas to offset their emissions. This is social inequality!" In addition, if "everything is reduced to data and the market," "will initiatives to reduce non-market GHG emissions (e.g. a project for people living in neighbourhoods that offered active mobility, in urban agriculture) become invisible and therefore be discriminated against?" Lastly, "Lifestyles differ around the world, diets vary (e.g. vegan, religious). Is there a risk that some lifestyles will be favoured and others will be discriminated against?" "Will this create culture-based discrimination?"

DEMOCRACY AND GOVERNANCE PRINCIPLE

"Who will regulate this system? The United Nations? The rich countries? How will abuses be monitored?" "Should an authority be established to regulate carbon footprints?" "If some CO_2 is saved, could those savings be transferred to family and friends?" "Should the recommendations that will be given priority be subject to public debate?"

The participants then engaged in several in-depth discussions, going back to their initial ideas to generate more ideas. Then, following close to 45 minutes of discussion, they selected their priority groups of ethical issues for 2025 by applying coloured labels. Two principles of the Montréal Declaration emerged: Autonomy, tied to freedom of choice; and Justice, which they associated with the Equity principle.

Table 3: Environment. First Deliberation Period: Formulation of Ethical Issues in 2025

Ethical issues in 2025	**1**	**2**
Description	Risk of algorithmic confinement due to this new prescriptive power and the configuration individuals' preferred spaces. How can individual and societal autonomy be maintained? How can we account for an initiative to reduce carbon emissions that is outside the system?	Carbon offsetting could favour the richest members of society. What limits should be assigned to them? Conversely, could people who consume only small amounts of carbon redistribute the carbon they saved? What about relations between Northern and Southern countries? Is there a risk of cultural discrimination?
Related principle	Autonomy and freedom of choice	Justice and equity

This selection of priority issues by the team led them to develop and give more clarity to two ethical issues in AIS. The first is related to the Autonomy principle, with potential actions to reduce non-market greenhouse gas emissions (e.g. a citizen initiative on daily mobility). The second issue concerns the Justice principle, with potential carbon offsetting for the richest members of society, or emission sharing for citizens who consume less than their limit.

SECOND DELIBERATION PERIOD: PROPOSALS FOR MANAGING AI FOR 2018-2020

In response to these issues, the team continued their discussions, brainstorming on the four related principles. The participants formulated several proposals for managing AI. Three of them are presented here, illustrating how the ideas were developed into a headline and lead for a front-page newspaper article.

Table 4: Environment, Second Deliberation Period: AI Management Proposals for 2018-2020

Management proposals in 2018-2020	1	2	3
Description	• Develop a code of ethics for system designers, programmers and managers (e.g. to ensure that the prescriptions support equality).	• Create a consumer ombudsman, an independent administrative authority, that is audited by the national democratic assembly. • Audits of the system, of the diversity of choices and recommendations, and the publication of transparent reports. • Support citizens in their autonomy.	• Provide substantive financial support to help people with the most modest means to adjust. • Allow people to exceed the annual target, but with a growing marginal cost for each additional tonne of carbon.
Instrument categories	Laws and regulations Code of ethics	Institutional actor	Incentives and support measures

These proposals, which reflect true institutional creativity (that went well beyond the examples of very general tools provided in the participant's booklet), are in keeping with the issues identified in the previous step. The proposal to create a consumer ombudsman to regularly evaluate the system by performing audits, being publicly accountable, and organizing citizen support also shows a further development of the ideas formulated in the previous step. It is on the basis of this proposal that the participants developed their headline newspaper article in the following step.

THIRD DELIBERATION PERIOD: WRITING A HEADLINE AND LEAD FOR A FRONT-PAGE NEWSPAPER ARTICLE FOR 2020

These measures were then expressed in a poster as follows. The team developed the following headline and lead for a newspaper article dated October 9, 2020:

Major Achievement for ConsumAI! One million subscribers in a week.

General Regulation on Opening Data for the Environment.

Following the passage of GRODE (the General Regulation on Opening Data for the Environment), which required personal data to be made public, the CONSUM'AI organization conducted a major survey on the freedom of choice of ECOFIT users and found many limitations and algorithmic bubbles. The first recommendation in the CONSUM'AI report is to develop the means for counter-expertise, and train of users to ensure true pluralism and everyone's participation in reducing greenhouse gas emissions.

So if the use of AI presents a certain potential for managing the environmental issues associated with consumer behaviour, this perspective also raises many ethical issues that must be properly framed.

2.3.

ADDRESSING ISSUES OF DIGITAL TRANSFORMATION IN THE WORKPLACE

Summary of the initial scenario:

Mining HR data to optimize work atmosphere

Peter has finally landed a job in a good law firm. After his first three weeks on the job, he meets with Marco from Human Resources for some personal mentoring. Marco explains that, going forward, the firm will be using AmbAI+, a conversational analysis AI that studies employees' attitudes and helps maintain a peaceful and productive atmosphere at work (the system analyzes all e-mails, telephone calls and conversations in work meetings). AmbIA+ provides personalized assistance, advice and training, but no disciplinary action is taken. Peter is told that all his conversations in the office have been monitored, except for those on October 15 and 16. "On several occasions you interrupted your co-workers in meetings to repeat the same ideas, and this has created some tensions. Apparently, the algorithm also detected periods of inactivity on the network, periods lasting several hours, when you had engaged in no conversation with your colleagues. This does not pose a problem in and of itself, but it is better to stay in touch with your team. Do you remember why you were inactive on the network at these times?" This not only has Peter worried, it also embarrasses him, and he wonders how these issues can be relevant.

The objective of this scenario was to open a discussion about the ethical issues related to businesses using AI to monitor and manage their employees. The AmbAI+ system imagined here is used to optimize performance and control the work atmosphere using data mining techniques.

The deliberations presented here are the result of a full day of discussions among a group of 10 engineers, AI designers, digital strategy managers, investigators, students and academics. Based on this scenario for 2025, the discussions led writing a headline and lead for a front-page article in the AI newspaper dated October 9, 2025: "First Employee Terminated Because of AI."

FIRST DELIBERATION PERIOD: FORMULATION OF ETHICAL AND SOCIAL ISSUES IN 2025

In this first part, the participants identified five categories of issues related to the development of AI applications in the workplace.

AUTONOMY

First, the discussions underscored the issue of respect for autonomy (in particular as it relates to employees' ability to act). The participants criticized a kind of manipulation of "how people feel things," and a "forced" organizational culture. They were troubled by how information was being saved, such as information on employees' behaviour and interactions with each other, for the purposes of cultivating an organizational culture. In operating this way, the company is somehow trying to standardize employees, which could lead to a great deal of tension (even "totalitarianism," if everything began to be measured in order for the business to exercise control over its employees), through insistent recommendations issued by the AI. Respect for autonomy is tied to respect for employees exercising a certain "free will" as well as respect for their emotions, which in this case has not been given due consideration, according to the citizens.

Should all this data be kept (the smallest actions are being observed) and used for this purpose? The participants therefore raised an issue related to "surveillance," which could limit the scope of employees' actions and speech (this is closely related to the issue of respect for privacy).

> "AI is watching"
> - A participant

The citizens put forward the need to foster autonomy by allowing everyone to work in whatever way is best for them, in the best interests of the business. Employees should be able to have control over their data, the employer should tell employees what data is being collected and use this data in a carefully circumscribed manner, and everyone should be free to "disconnect," especially to maintain a boundary between what is professional and what is private.

PRIVACY

The participants debated where this boundary between the private and the professional in a business is situated, and concluded that it is difficult to define. Some of them felt that the AIS in question would intrude on employees' privacy, based on their conversations:

> "When should discussions be considered personal?"
> - A participant

On the other hand, some participants mentioned that, ordinarily, anything related to one's private life should not be discussed in e-mails or telephone calls using company equipment (meaning that it would not be analyzed by the AIS presented in this scenario). These participants asked: Is there a place for one's private life in a business?

A consensus nevertheless emerged about the use of AI: as a business tool, it must never, under any circumstances, be used to analyze anything private. The issue then becomes how this boundary should be defined, in order to clearly identify when AI can and cannot be used (i.e. there is a need to define what data is purely work-related and can therefore be used in the system's analyses).

The behavioural analysis performed by the AIS was also criticized. It could infringe on both privacy and autonomy (meaning that the citizens then questioned the ethics of using an AIS to track conversations and behaviour, even if they were work-related). Here they criticized a form of intrusion and breach of confidentiality that would result from this constant surveillance:

> "It's as if Big Brother is watching."
> - A participant

Some of the participants noted that these issues were not specific to AI, while others believe that the high level of traceability afforded by AI enhances the relevance of this issue.

WELL-BEING

The participants felt that this type of system could have both positive and negative impacts on employee well-being. While this technology can be used to help employees and improve the quality of their work life (by helping improve relationships or revealing incidents of harassment or intimidation, or even by helping prevent suicide), it appears that the technology can also cause harm ("destabilizing" employees, making them "distraught," standardizing employee behaviour). The participants agreed that the scenario asks too much of employees, who should not be obliged to justify all of their actions. Here the citizens highlight a dilemma between employee well-being and freedom (just how far can a business go in monitoring employees' actions in the interests of protecting their well-being without unduly restricting their freedom to act? When can surveillance be considered "well used"?). The citizens also found a correlation between well-being and the performance objectives: a happy employee is also more productive. So protecting an employee's well-being appears to be good for both the business and the individual.

TRANSPARENCY

The citizens began by discussing the lack of respect for employee consent, since the employee did not know that he was "under surveillance," and they called for more transparency from the employer, who should have informed him about what was and was not being recorded (in particular, through the business's employee training).

> "Employee consent is important in data collection. Transparency with employees is essential."
> - A participant

This transparency issue raised several questions: What are the rules on relationships within a business? What hierarchy has been established for the importance of the data collected? Who has access to it? Do employees have the right to see their boss's data?

The transparency issue also refers to AI or how it can be made interpretable. Here the citizens spoke of the need to communicate about how the algorithm works, including so that individuals will "buy in" to these new systems. They also mentioned the importance of not making a decision based on the conclusion of an AI technology that cannot be explained.

PRODUCTIVITY AND PERFORMANCE

The issue of employee performance and how it is evaluated was then raised several different times. To what extent does AI need to intervene in the interests of improving employee productivity (e.g. by stopping employees from repeating themselves in group discussions)? Is this truly important? Here it would be necessary to define exactly how the AIS can measure and improve performance. There is a risk of emphasizing productivity at the employee's expense, at the expense of his or her personal development and behaviour in the workplace. Does an IA application like this truly give employees an opportunity to improve? Should expectations and objectives be tailored to the individual?

"AI doesn't forget anyone."
- A participant

On the other hand, the participants were concerned with the very definition of what should be considered productive (and counter-productive), as essential to the debate. Which indicators should be used? How are they relevant? Who can and should determine them? Do they need to be related to the business's objectives? Should these objectives be reviewed, based on the AIS's analyses? Even though the system is being used to confirm that employees are "performing," by interpreting the employees' results the AIS could just as well impede innovation, in particular since certain tasks are easier to measure than others.

All the discussions about defining performance, productivity and respect for employee well-being led to the conclusion that these issues are closely tied to corporate culture, which can vary widely from one business to the next and reflect various objectives, interests and values.

Some citizens around the table were concerned that the adoption of AI-based tools is pushing companies to impose notions of performance that are systematically associated with a score, which could result in a form of standardization. Others pointed out that these practices already exist, and will only be amplified or even "industrialized" by AIS, which can process more information, much more quickly. Others mentioned that it may nevertheless be useful to standardize practices, in particular as a response to business needs.

A debate then ensued on the so-called objectivity of AI, raising questions such as: How can we know the decision-making and automation criteria? How will AI define an absence of employee activity? How can we balance performance, objectivity and standardization by AI technologies, on the one hand, with human subjectivity, specificity and arbitrage, on the other?

After discussing these issues for more than an hour, the participants selected three of the five issues as being, for all intents and purposes, priorities for 2025. In some ways these issues reflect principles set forth in the preliminary version of the Declaration.

Table 5: Priority Issues

Ethical issues in 2025	**1**	**2**	**3**
Description	How can we introduce a performance measurement framework while respecting both the goals of the company (productivity) and individual (normalization)?	How can we ensure that companies and employees understand AI (and ensure their buy-in)?	How can AI ensure (contribute to, support) employee well-being?
Related principles	Performance	Transparency (knowledge)	Well-being

SECOND DELIBERATION PERIOD: PROPOSAL FOR MANAGING AI, 2018-2020

For this second part of the activity, participants were asked to formulate recommendations and imagine what kinds of solutions could be implemented in response to these three issues. For each of the issues identified, the participants formulated a wide range of more or less restrictive mechanisms, and ultimately selected six principal mechanisms to implement that would cover all the issues.

In response to the **performance** issue, the participants first recommended organizing **continuous training activities** in businesses to support people in each stage of their company's "digital transformation." Digital **education** was proposed to encourage the establishment of a learning climate, but also to reduce the fear that can come with implementing AI in the workplace.

> "Continuous training for all employees at each stage of a business's digital transformation (encourage a continuous learning climate)."
> - A participant

In addition, participants proposed establishing an **independent administrative authority (IAA)** and a **correspondent** in the company as potential solutions, in particular in order to guarantee respect for the GDPR[2] (which should be extended to the traceability of data and the explainability of algorithmic decisions). The correspondent would be charged with applying the rules and supporting the complainant when a problem arises, and could seek the assistance of the IAA if required. The participants also proposed creating **indicators** that are directly tied to not only the business's objectives (e.g. financial results) but also to certain "human values" (e.g. the employee's well-being). To this end, the citizens felt that it is absolutely essential to pass a **law**, or else this score would be solely correlated with the business's financial interests.

A recommendation was also made for the government to create a **public research program** on algorithm interpretability and transparency (to close the gap in private research on these issues). The goal would be to understand how algorithms make decisions and limit the monopoly held by large AI companies. For the same reasons, **"pilot groups"** (or "test groups") should be established in order to measure the impacts (including the psychological and sociological impacts) of businesses using AI and confirm its relevance and usefulness.

In response to the ***transparency*** issue, some participants argued for less restrictive measures, such as a **mandatory communication** on the various rules followed, the salary data used, the objectives behind their collection (individual or group scores?), and what can be deduced from the data or the results of the analyses performed by the pilot groups mentioned above. This communication must be addressed to all departments (HR, IT, Marketing, Legal Services) and include AI concepts, with training on what an algorithm is and how it learns from data.

To protect ***well-being,*** the participants recommended **an annual evaluation** of employees' perceptions of the use of AI applications. This evaluation could eventually be consolidated by a committee (such as an occupational health and safety committee) that would be responsible for responding when problems arise. The committee should foster good relations between workers and the employer and ensure that everyone's way of working is respected, even when the technology is applied. A **certification** (or label) should be developed, guaranteeing good ethical, environmental and societal practices, and it should be imposed by the state. This will guarantee that businesses meet minimum criteria in order to optimize productivity and performance. **Indicators** of well-being need to be taken into account, just like indicators of performance.

In response to the well-being and transparency issues, the citizens proposed that a law be introduced to define and impose interpretability of AIS (in particular, justifying the decision and guaranteeing access to explicit rules), and it should set minimum criteria for protecting individual well-being (including a right to regularly disconnect)

[2] General Data Protection Regulation, the European regulation that took effect on May 25, 2018.

in order to guarantee the protection of fundamental rights that could be threatened by the development of AI.

For all these issues and in the interests of "regulating without penalizing," an official **charter**[3] would be created, covering the rights, duties and values that should be defended to protect individuals in a company.

In the end, the citizens reached a consensus on **6 recommendations** that covered the major points of the previous proposals:

Table 6: Proposals Retained

Management proposals	**1**	**2**	**3**	**4**	**5**	**6**
Description	Law that defines and imposes AI interpretability and establishes minimum criteria to protect individual well-being	Certification (or label)	Training for different company stakeholders	Updating the labour code to reflect the new digital reality	Creating a charter of rights and duties	Funding public research and pilot studies on AI and its impact on the workplace
Related issues	*The six (6) recommendations were formulated in response to the three (3) priority issues*					

[3] The participants pointed out that a charter is not as limiting in France as it is in Quebec.

THIRD DELIBERATION PERIOD: WRITING A HEADLINE AND LEAD FOR A FRONT-PAGE NEWSPAPER ARTICLE IN 2020

This step involved storyboarding one of the solutions proposed in 2020. Here the participants outlined the risks of using AI in businesses and one of the planned measures for addressing these risks.

First Employee Terminated Because of AI

An employee was terminated after three weeks of work on a recommendation made by an AI application. The employee appealed his dismissal with the labour board and a decision was handed down following a debate in the legislative assembly. A law that will provide a legal framework for this practice will be put to a vote.

The participants mentioned that in their discussions they paid particular attention to the potentially *negative* impacts of AI in the workplace. However, they recognized that there could also be many benefits to using AI, for both employees and businesses, and that these benefits could be addressed in another forum. Introducing an internal body supported by a legislative mechanism would appear to be indispensable to responsible AI in private-sector organizations.

3. DISCUSSION OF THE THEME OF CULTURE with members of the Coalition for the Diversity of Cultural Expressions (CDCE)

In order to address issues related to AI developments in arts and culture, a discussion workshop was organized with the Coalition for the Diversity of Cultural Expressions (CDCE) on September 25, 2018, bringing together 11 experts and stakeholders in arts and culture. A series of discussions were organized around three themes:

1. copyright,
2. cultural diversity, and
3. propaganda and manipulation.

Following the discussions, the CDCE produced a particularly relevant brief[4] describing various challenges and opportunities related to AI developments in the field of culture. This brief also presents the ethical principles essential to responsible AI in culture and the main recommendations that emerged from the discussions on September 25. This section summarizes the discussions of September 25 and the main points of the CDCE's brief.

3.1.

THREE THEMES PROPOSED BY THE DECLARATION TEAM TO FACILITATE DEBATE ON AI DEVELOPMENT ISSUES IN THE FIELD OF CULTURE

COPYRIGHT IN A CONTEXT OF CO-CREATION BY AIS

The use of AIS to generate works at very low prices (e.g. by generative adversarial networks) in music, in visual arts, TV series or in writing newspaper articles will raise the copyright issue in a novel way. Should the copyright belong to the writers of the examples from which the algorithms learn, or to the programmers of the algorithm, or even proponents of the project? Does the remix produced by a generative algorithm constitute plagiarism? If AI replaces artists, what impact will this have on cultural diversity? And what if an AI application "interprets" a work of art? What access should algorithms be given to our artistic heritage?

This problem is of particular concern in the context of AIS that generate creations ("Applications of AI in the cultural field," CDCE brief, p. 3).

ISSUE OF CULTURAL DIVERSITY WHEN NEW AI APPLICATIONS PRODUCE RECOMMENDATIONS BY ALGORITHM

Given the risks of standardized tastes and behaviours, of an "algorithmic bubble," of recommendation algorithms capturing our attention and formatting our choices, how can we maintain a diverse cultural offering? What sort of free, autonomous and critical reception practices should users be encouraged to use? How can we disconnect (connected objects, domestic robots), given the strategies for attention capture? How will the algorithm's contributions be made transparent and be explained to users? Should a public policy on culture be proposed for this diversity issue? What

[4] For more information on issues related to the development of AI and cultural diversity, see: https://cdec-cdce.org/en/ethical-principles-for-the-development-of-artificial-intelligence-based-on-the-diversity-of-cultural-expressions/

will be the new funding mechanisms for cultural diversity?

Above all, this problem concerns algorithms for data recommendations and use (see "Applications of AI in the cultural field," CDCE brief, p. 3).

ALGORITHMIC CENSORSHIP AND CONTEMPORARY ART

Recognition algorithms are used by social media to exercise censorship in ways that are sometimes considered excessive. Contemporary artists have responded to these applications with interventions intended to both make these rules visible and bypass them. How much critical freedom will tomorrow's contemporary artists have? What kind of training do artists need in order to develop critical knowledge?

Above all, this problem concerns recommendation algorithms and data use (see "Applications of AI in the cultural field," CDCE brief, p. 3).

3.2

PROMOTING CULTURAL DIVERSITY IN THE AI ERA

DISCOVERABILITY AND HOMOGENEITY OF CULTURAL CONTENT

"When recommendations are based, among other things, on the popularity of the content, they contribute significantly to the concentration of listening (0.7% of titles represent 87% of plays on online music services in Canada), favouring a minority of artists."
(CDCE brief, p. 5).

Discoverability was identified as an important issue in this new era of AI in cultural production. If the parameters of AIS are correctly set, they may become tools for cultural diversity by expanding global audiences. They could offer relatively diverse content by, for example, allowing artists to increase their visibility without necessarily having to be supported by intermediaries, even without production costs. On the other hand, the participants were concerned about the risk that cultural content would be standardized and about real net neutrality. For example, francophone content from Quebec is rarely proposed by recommendation algorithms on cultural platforms such as Netflix, Amazon and Spotify. These platforms carry massive content offerings, so much that users often just rely on the recommendation algorithm to make choices for them. The participants mentioned the risk that the algorithms will confine individuals to "a particular taste," preventing them from discovering any other content than what is recommended, based on their previous selections. The risk is that this will lead to homogeneous cultural content, in particular because most artists adapt their creations to this mode of dissemination.

The participants said that the audiovisual sector needs to make an effort to make their works discoverable on digital platforms. The problem

does not only stem from AIS; it is also an issue tied to the industry's objectives. If governments have a responsibility for ensuring cultural diversity, multinationals challenge this power. For example, Canada was the first signatory to the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions. Even though some content may be funded from public sources, the dominant model remains a business model, which may be a problem, such as when we see that Quebec books are not recommended by the Amazon algorithm (on Amazon.ca), despite a desire for them in the province.

The participants therefore insisted on a democratization of creative power and on using AIS for this purpose, such that they become significant agents of the cultural ecosystem for the emergence of creations.

CONSEQUENCES ON EMPLOYMENT

"In May 2018, researchers released the results of a survey of more than 350 AI researchers. On average, they predict that AI will be able to outperform humans to produce school-level essays in 2026, popular songs in 2028 and best sellers in 2049."
(CDCE brief, p. 4)

The possibility that AI will be used to generate an entire work without any involvement by an artist gives some cause for concern. By whom will these works be developed? Only by businesses with a certain amount of capital? Is this compatible with the development of a society consisting of more artists and more diversity? On this point the participants criticized the risk that there will be a closed loop (in particular in light of work on the European directive on this subject), in which the work of only a tiny minority of artists will be favoured by the algorithms. The participants were afraid that the number of artists will decline precipitously.

How can human artists set themselves apart from AI applications? How will this affect personal and collective capacities to create and innovate? In an experiment conducted by ACTRA (the Alliance of Canadian Cinema, Television and Radio Artists), subjects were unable to distinguish between characters created by an AI application and those created by people. This appears to go well beyond the issues raised by deepfake, since in this case new faces were created. But this is a role played by many of ACTRA's members, who work in the video game industry. The issue here is the job losses that these technologies could engender.

The issue of remuneration was also raised, opening the door to a discussion of copyright. Who will be remunerated, and how, if algorithms are used to create works? The remuneration of actors is calculated in number of days and by residual rights. In terms of residual rights, compromises are to be expected, such as standards governing the re-use of works. The participants also wanted to soften the potential impact of AI on artist recognition, which could be less than anticipated, in much the same way that the digital book did not kill the paper book.

RETHINKING COPYRIGHT

"Obviously, the dematerialization of cultural content, technological changes and the arrival of new players who have transformed business models have a major impact on artists' remuneration and the payment of copyright royalties."
(CDCE brief, p. 6)

Various questions were raised on this subject: Where will the royalties go? Which regulations will govern copyright? How much of a manuscript does an algorithm need to produce another one? How will AI be able to draw from other books to bring new ones to market? The arrival of creative AIS therefore raises major property and cultural content issues. Concerning music, the high processing capabilities and greater amount of available data, combined with major advances in algorithm performance,

has led to the creation of artistic "generation" tools. In music, generally, learning is already based on what was done before. This is what AIS do, but at unprecedented speeds.

The participants recognize that AI is already flouting copyright laws, and they questioned whether this type of right could be granted to a machine. Given that satire and parody are already considered exceptions to the rules on copyright, would it be possible to allow an exception for AI? The Copyright Act is currently under review. In this case the participants felt that it is essential for the law in its new form to include amendments related to AIS developments.

3.3

THE CDCE'S ETHICAL PRINCIPLES

In its brief, the CDCE recommends adopting four ethical principles to guide the development of AI and prevent abuses in the cultural field. According to the CDCE, application of these principles should ensure that AI will more successfully integrate cultural issues in general, and diversity issues in particular. These three principles are consistent with those developed in the final version of the Montréal Declaration, but integrate priorities specific to culture and the arts.

The CDCE put forward a principle on the diversity of cultural expressions, which directly reflects the principle of inclusion and diversity. However, in this case the CDCE specifies that this principle should ensure that AIS:

"- enhance local cultural and linguistic content within the populations from which it originates, thus promoting social cohesion as well as the local economic fabric;

- encourage users to make discoveries outside their environment;

- facilitate the transition between technological families (e.g. Apple), rather than locking them in;

- promote interaction and content sharing."
(CDCE brief, p. 7)

The CDCE also proposes a principle on enhancing culture, artists, creators and producers of cultural content, meaning that AIS should help avoid the current devaluation of cultural content and be prevented from "promoting excessive appropriation of revenues that should be directed to cultural ecosystems" (CDCE brief, p. 7). While this principle is related to other principles of the Declaration, above all it is a call to respect the sixth equity principle: "The development and use of AIS must contribute to the creation of a just and equitable society," and the related sub-principles.

The CDCE then proposes a transparency and dialogue principle (transparency in terms of the algorithm's code but also the data used, and dialogue with users, in particular). This principle calls for respecting the fifth principle of the Declaration regarding democratic participation: "AIS must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control"; and the second principle of the Declaration, on autonomy: "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."

Lastly, the CDCE proposes a principle on the primacy of the public interest, which it defines as follows: "Not all technological innovations are desirable. The development of AI should always focus on improving the quality of life of the population, social cohesion and democratic practices. Governments must defend the public interest against developments that could have rather negative impacts on society." (CDCE brief, p. 8). Respect for this principle is aligned with respect for first principle of the Declaration, on well-being: "The development and use of artificial intelligence systems (AIS) must permit the growth

of the well-being of all sentient beings," but also the eighth principle of the Declaration, on prudence: "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."

The CDCE's ethical principles	Principles of the Montréal Declaration
Diversity of cultural expressions	Principle 7, Diversity inclusion principle
Enhancement of culture, artists, creators and producers of cultural content principle	Principle 6, Equity principle
Transparency and dialogue principle	Principle 2, Respect for autonomy Principle 5, Democratic participation
Primacy of the public interest principle	Principle 1, Sustainable well-being Principle 8, Prudence

3.5

SELECTED RECOMMENDATIONS

Various recommendations emerged from the September 25 discussions. They were formulated with an eye to promoting Quebec cultural content and making citizens aware of the impacts that AI development can have on culture. First, to foster diverse cultural expression in the digital world, the participants recommend that minimal requirements be set on the representation of Canadian cultural content in the recommendations made by algorithms. This is already the case for Quebec TV and radio. The participants do not believe that free markets will develop these requirements on their own, so they must be formulated in laws and regulations.

During their discussions the participants recognized that literacy in AI development is essential. People need to be equipped to understand where the recommendations made by algorithms will lead them. The participants recommend implementing a user education policy, intended to counter the false impression of choice by encouraging users to vary their browsing and stay vigilant to the influence exercised by algorithms. This policy will take the form of education in how to exercise critical choice. Everyone should develop a form of intellectual self-defence, beginning in childhood. The participants also recommend raising awareness among IT developers of AIS's impact on culture.

These discussions also produced a recommendation concerning the transparency and explainability of algorithmic recommendations. Users should be systematically informed when a recommendation has been made by an AIS, and they should have easy access to explanatory information on both how algorithms work and the existence of other cultural content.

The participants also recommended that the businesses developing AIS that have an impact on culture spend a portion of their sales revenue on promoting cultural diversity, such as by funding certain libraries, cultural events or media. The participants also support the monitoring of taste profiling and the protection of personal information, or even the development of an "AI in culture laboratory" to observe algorithms, learn how to interact with them and, eventually, how to influence their development.

Two of the main recommendations that emerged from these discussions were further developed in the CDCE brief:

1. education and training, and
2. revisions of laws affecting the cultural community.

These recommendations are consistent with those formulated for other sectors during last winter's co-construction: 26% of the recommendations refer to legal provisions and 19% to training *(see Part 3 Summary report of the recommendations from the winter co-construction workshops).*

4. BRIDGING THE GAP BETWEEN THE PUBLIC CONSULTATIONS AND A NEW GENERATION OF RESEARCHERS: POLICY BRIEF SIMULATION

4.1.

DESCRIPTION OF THE ACTIVITY

To bridge the gap between the emerging generation of researchers and citizens, the Declaration organized a simulation in partnership with the Comité intersectoriel étudiant (CIÉ) of the Fonds de recherche du Québec (FRQ) and the École de politique appliquée (EPA) at the Université de Sherbrooke. The simulation was held in association with the Journées de la relève en recherche (J2R) organized by ACFAS. The purpose of the "Policies and Artificial Intelligence" simulation was to bring together students representing a new generation of young researchers to produce three policy briefs on AI. The objective was to allow this new generation to take part in the discussions on AI and the ethical and social issues around its development. CIÉ members were united in their response to this theme:

"AI was selected because it has cross-sectoral dimensions and encompasses issues of particular interest, since they blend science, society and the development of public policies. In this simulation, AI allowed members of the emerging generation of researchers to take part in the discussions and think about Quebec's leadership position in this area."
[translation] (participants guide, p. 5).

With this in mind, the Montréal Declaration provided three problems that had been identified during the citizen co-construction activity in the winter of 2018. We felt that it was relevant, both as part of work on the Declaration and for the work of the young researchers selected for this activity, to discuss these themes and issue recommendations. These three problems highlight particularly sensitive issues in AI development that urgently need to be debated:

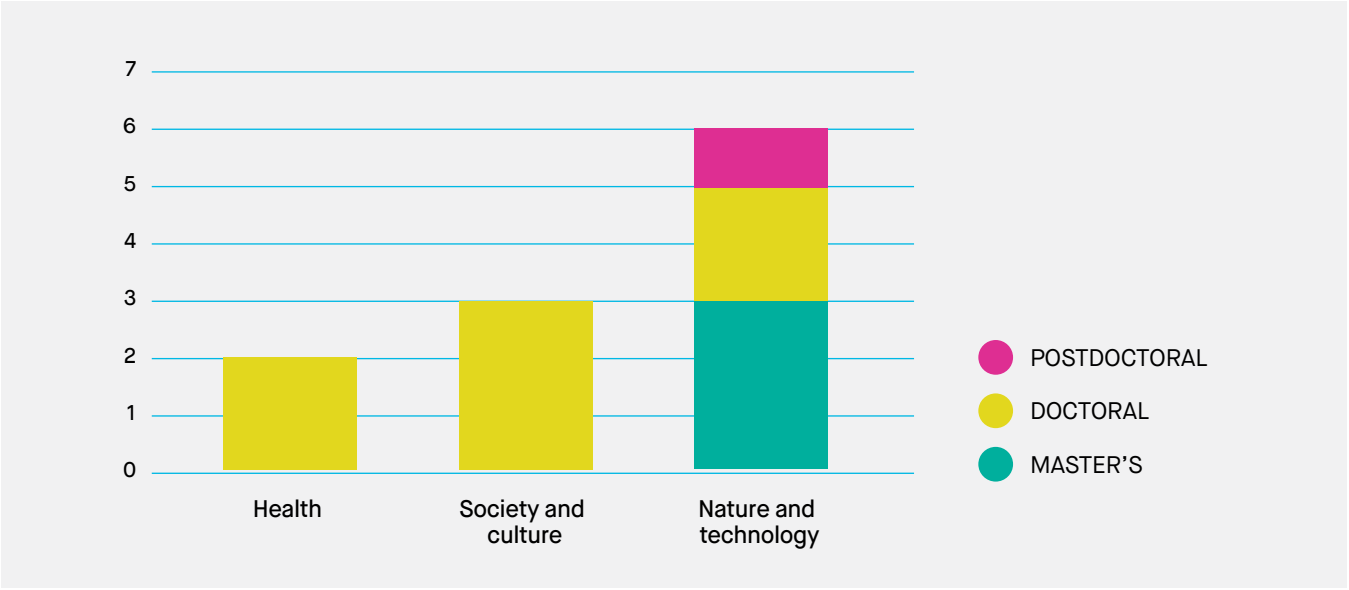
1. The security and integrity of AIS, *i.e. How can we maximize the positive impacts while minimizing the adverse effects of AI development?*
2. AI, the media and the manipulation of information, *i.e. How can we fight the dissemination and amplification of fake news and disinformation campaigns? How can we foster a democratization of access to information while encouraging critical thought and informed decision making?*
3. Public, private and participative governance: the digital commons. *Which of these types of governance is the most appropriate? What checks and balances are needed?*

The simulation had three objectives: "(1) to familiarize the participants with writing and presenting a policy brief, (2) to facilitate the acquisition of skills related to developing science policies, and (3) to analyze a social problem from a scientific point of view." [translation] (participants guide, p. 6). Policy briefs were developed in an exercise that was designed, first and foremost, to serve pedagogical purposes; the objective was not to disseminate the results to decision-makers and stakeholders. However, the recommendations in these briefs reveal the views of students from the emerging generation of researchers, and they are particularly relevant. The briefs have therefore been

attached to this report, even though they do not contain actual recommendations on public policies. We decided to present the briefs in their original form, unrevised, so as not to misrepresent their work, and in an effort to present their contribution as authentically as possible (see Appendix 2).

The activity was carried out on October 18 and 19, 2018 at the Université de Sherbrooke and involved 11 students with varying levels of education and different types of expertise.

Chart 5: Profile of Participating Students, Based on Area of Study (according to the three FRQ funding areas)



The students were assigned to three groups, with each group addressing one of the problems and led by a facilitator (someone who was independent of the Declaration, to avoid influencing the recommendations in any way). The students were given four presentations on AI and on writing policy briefs. Then they were given only six hours to draft their briefs and prepare an oral presentation on their work. The three groups presented their briefs to a jury[5] on the morning of October 19. The contest was won by Team 2 (which had been given the problem "AI, the media and the manipulation of information").

[5] The jury was consisted of three members:

Claude Asselin, Full Professor, Department of Anatomy and Cellular Biology, Faculty of Medicine and Health Sciences, Université de Sherbrooke [representing ACFAS].

Benoit Sévigny, Director of Communications and Knowledge Mobilization [representing FRQ]

Nathalie Voarino, Doctoral candidate in Bioethics, Scientific Coordinator at the Montréal Declaration [representing the Declaration]

4.2

PROBLEMS IDENTIFIED ON THE BASIS OF CITIZEN CONCERNS

Problem 1. Public Security and System Integrity

During the consultations, the citizens acknowledged that the development of AI could help make our physical and digital environments safer. For example, in an intelligent city, intelligent transportation systems may reduce traffic accident rates; in public health, epidemiological models may allow authorities to better predict the spread of illnesses; and in cybersecurity, IT security specialists are using AI to recognize attacks.

However, the citizens also recognized that certain conditions are necessary in order to ensure that AI advances are beneficial to public security. Guaranteeing "proper" use of AI through system integrity and security is fundamental to the responsible development of these technologies.

AI's negative impacts on public security can take four different forms:

1. **An AIS designed to threaten public security.**[6] For example, the use of AI for cybercrime (identity theft, hacking into nuclear power stations, etc.), political destabilization (targeted propaganda, the creation of fake videos, etc.) and the automation of military equipment (drones, robot soldiers, etc.).[7]
2. **An AIS that uses information for purposes other than those originally intended.** In this case, the citizens fear that complete medical records could be misused by insurance companies, that school records could be used to automate the labour market, or that automated traffic systems could be used to follow and monitor road users.
3. **Willful hijacking of AIS.** A person with malicious intent could directly target how the algorithm works[8], such as by outwitting a facial recognition system to gain access to protected data. Someone could also take advantage of the security challenges created by the proliferation of connected objects[9], such as to take control of an autonomous vehicle, or to paralyze a network with a massive denial-of-service attack.[10]
4. **An AIS that has been poorly evaluated:** An AIS whose reliability or robustness has been overestimated and which has caused an accident.[11] For example, the citizens mentioned that an accident involving an autonomous truck or a systematic error by a medical diagnosis program can have serious consequences.

The citizens therefore wondered how to limit the negative impacts of AI on (public) security. They raised various potential dilemmas related to system security and integrity:

> Could respect for transparency (a frequently mentioned imperative) jeopardize security by facilitating hacking?

> Does providing the most security possible necessarily mean that the system will be less

[6] Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. 2018; (February 2018). Available at: http://arxiv.org/abs/1802.07228

[7] Bouvet M, Chiva E. *Un regard (décalé ?) sur Intelligence Artificielle et Défense/Sécurité* - CGE [Internet]. Conférence des Grandes Écoles. 2016 [visited on Sept. 3, 2018]. Available at: http://www.cge.asso.fr/liste-actualites/un-regard-decale-sur-intelligence-artificielle-et-defensesecurite/

[8] Kurakin A, Goodfellow I, Bengio S, Dong Y, Liao F, Liang M, et al. *Adversarial Attacks and Defences Competition* [Internet]. 2018 [visited on Sept. 3, 2018]. Available at: https://arxiv.org/pdf/1804.00097.pdf

[9] Zhang Z-K, Cho MCY, Wang C-W, Hsu C-W, Chen C-K, Shieh S. *IoT Security: Ongoing Challenges and Research Opportunities.* In: 2014 IEEE 7th International Conference on Service-Oriented Computing and Applications [Internet]. IEEE; 2014 [visited on Sept. 3, 2018]. p. 230–4. Available at: http://ieeexplore.ieee.org/document/6978614/

[10] Franceschi-Bicchierai Lorenzo. *How 1.5 Million Connected Cameras Were Hijacked to Make an Unprecedented Botnet* [Internet]. Vice Motherboard. 2016 [visited on Sept. 3, 2018]. Available at: https://motherboard.vice.com/en_us/article/8q8dab/15-million-connected-cameras-ddos-botnet-brian-krebs

[11] Amodei D, Olah C, Brain G, Steinhardt J, Christiano P, Schulman J, et al. *Concrete Problems in AI Safety* [Internet]. [visited on Sept. 3, 2018]. Available at: http://arxiv.org/abs/1606.06565.pdf

efficient (it must be secure without becoming inoperative)?

> And, more generally, how can the positive impacts of AI development be maximized while preventing the adverse effects?

Other relevant references:
Asilomar AI principles
Adversarial ML

Problem 2. AI, the Media and the Manipulation of Information

The citizens were concerned about the risk that users may be manipulated, to the extent that their actions are increasingly affected by the AI mechanisms influencing their decision-making, often without their knowledge or through incentives. This raises a problem of trust in these applications, since there is a form of interference with one's autonomy, and a risk that the systems will give direction to actions (for example, based on private interests). For example, the citizens wondered whether new technologies derived from AI could create a new lobbying class, which could at times become too powerful. To maintain a certain level of freedom in the choices suggested by the AI and to avoid placing blind trust in these applications, it would therefore be important for all citizens and professionals interacting with the AI application to cultivate critical thinking skills.

Although propaganda is not a new phenomenon, it can now be created and disseminated through fake news and disinformation campaigns with unprecedented ease and speed. This includes through platforms for creating and disseminating content online (through social networks, blogs and Internet sites, and discussion forums) that is structured according to attention retention, advertising and recommendation models.[12,13,14,15] This phenomenon is also amplified by an ability to very accurately target individuals by collecting and analyzing personal data, as we saw in the Cambridge Analytica scandal[16]. This reduces the diversity of the content seen by each individual to the sum total of whatever is closest to what he or she has already liked, shared and commented on. This leaves people mainly exposed to ideas that they agree with, such that the individual is caught in a "filter bubble,"[17] raising doubts about the likelihood that any citizen today will develop critical thinking.

Each of the major social media companies has announced a series of measures to limit the propagandist potential of their tools (see the transparency reports of Facebook, Google and Twitter[18]), but is this enough? How can we ensure that these tools, which have democratized access to information and interpersonal connections, are not used to democratize propaganda? How can we fight the dissemination and amplification of fake news and disinformation campaigns, to save democracy? How can we foster the democratization of information access while encouraging critical thinking and informed decision-making?

[12] Ingram M. *Fake news is part of a bigger problem: automated propaganda* [Internet]. Columbia Journalism Review. 2018 [visited on Sept. 3, 2018]. Available at: https://www.cjr.org/analysis/algorithm-russia-facebook.php

[13] Lewis P. *"Fiction is outperforming reality": how YouTube's algorithm distorts truth.* The Guardian [Internet]. February 2, 2018 [visited on Sept. 3, 2018]; Available at: https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

[14] Marwick A, Lewis R. *Media Manipulation and Disinformation Online* [Internet]. Data & Society Research Institute; May 2017 [visited on Sept. 3, 2018]. Available at: https://datasociety.net/output/media-manipulation-and-disinfo-online/

[15] Tusikov N. *Regulate social media platforms before it's too late* [Internet]. The Conversation. 2017 [visited on Sept. 3 2018]. Available at: http://theconversation.com/regulate-social-media-platforms-before-its-too-late-86984

[16] Cadwalladr C, Graham-Harrison E. *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach.* The Guardian [Internet]. March 17, 2018 [visited on Sept. 3, 2018]; Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[17] Pariser E. *The filter bubble: what the Internet is hiding from you.* London: Penguin Books; 2012.

[18] Preliminary Facebook report: https://transparency.facebook.com/community-standards-enforcement/
Google Transparency Report: https://transparencyreport.google.com/about
Twitter Transparency Report: https://transparency.twitter.com/fr.html

Other relevant references

Caplan R, Hanson L, Donovan J. *Dead Reckoning, Navigating Content Moderation After Fake News.* Data & Society Research Institute; February 2018 [visited on Sept. 3, 2018]. Available at: https://datasociety.net/output/dead-reckoning/

Foisy P-V. *Facebook veut s'attaquer aux fausses nouvelles au Canada* [Internet]. Radio-Canada.ca. [visited on Sept. 3, 2018]. Available at: https://ici.radio-canada.ca/nouvelle/1109432/fake-news-facebook-fausses-nouvelles-canada-verification-faits

Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. *The science of fake news.* Science. March 9, 2018; 359(6380):1094-6.

Jeangène Vilmer J-B, Escorcia A, Guillaume, M, Herrera J. *Les manipulations de l'information, un défi pour nos démocraties* [Internet]. Paris, France: CAPS and IRSEM; August 2018. Available at: https://www.defense.gouv.fr/irsem/page-d-accueil/nos-evenements/lancement-du-rapport-conjoint-caps-irsem.-les-manipulations-de-l-information

Internet sites to visit

The Computational Propaganda Project: http://comprop.oii.ox.ac.uk

Observatory on Social Media: https://truthy.indiana.edu

Conversation AI: https://conversationai.github.io

"A Citizen's Guide to Fake News" Center for Information Technology & Society, UC Santa Barbara: http://cits.ucsb.edu/fake-news

Problem 3. Public, Private or Participative Governance: Digital Commons

The citizens often raised issues about how the management of AI development will be shared by public and private institutions, as well as the related risks, such as conflicts of interest, how to protect the independence of institutional actors and public institutions, the market value of data, and privacy protection.

The risk that a private monopoly will emerge in AI development management was also mentioned several times. Some participants expressed concern that monopolies would emerge, in particular since a few companies own massive amounts of data (which are needed to make AI work). Their market power is bolstered by mergers with new, smaller service providers[19].

With respect to governance by the state, a legal framework for AI comes with its own risks and challenges[20], such as being overly focused on application capabilities at the expense of protecting human values[21], such that one might wonder if it is even possible to regulate AI. This raises doubts about the real power of the state[22].

Although discussions of governance issues often place public institutions at odds with private companies, an alternative has been proposed: participative governance. This mode of governance places citizens directly in control, and may involve carrying out a major public consultation or creating a permanent consultation forum.

In the context of participative governance, the participants proposed letting users make a major contribution to the design and management of AI tools. This participation could take the form of design thinking, using open-source equipment. Such equipment, which is accessible to all, was associated with the concept of digital commons, i.e. all the shared and co-created resources and knowledge that are available free of charge (e.g. open-source software). This is more than just a form of ownership: it is a mode of cooperative organization that guarantees horizontality (exchanges between peers) and freedom of expression[23]. This type of organization depends on letting the actors themselves choose the forms of regulation.

"Digital deployment is characterised by Internet communities. This process

[19] *Big data: Bringing competition policy to the digital era* - OECD [Internet]. [cited 2018 Sep 3]. Available from: http://www.oecd.org/competition/big-data-bringing-competition-policy-to-the-digital-era.htm

[20] Scherer MU. *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies.* Harvard Journal of Law & Technology, Vol. 29, No. 2, Spring 2016. http://dx.doi.org/10.2139/ssrn.2609777

[21] Ambrose ML. *Regulating the loop: ironies of automation law.* 2014;38.

[22] Danaher J. *Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems* [Internet]. Philosophical Disquisitions. 2015 [cited 2018 Sep 3]. Available from: http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html

[23] Crosnier HL. *Communs numériques et communs de la connaissance. Introduction*. tic&société. May 31, 2018;(Vol. 12, N° 1):1–12.

presupposes the emergence of significantly new organizational forms supported by information technologies, in particular open-source movements, and Web 2.0."[24] [translation]

This mode of governance is not without its challenges, including the fact that it is vulnerable to various forms of enclosure (fewer common uses), instituted by the state as well as by companies[25].

These issues raise a number of questions: What is the best way for AI governance to be shared between public, private and participative management? Does tension necessarily exist between these different modes of management, and which is the most appropriate? Is it necessary to benchmark these types of management, and if so, which types of guidelines should be put in place?

Other relevant references

Chessen M. *Encoded laws, policies, and virtues: the offspring of artificial intelligence and public-policy...* [Internet]. Medium. 2017 [cited 2018 Sep 3]. Available from: https://medium.com/artificial-intelligence-policy-laws-and-ethics/encoded-laws-policies-and-virtues-the-offspring-of-artificial-intelligence-and-public-policy-3dfb357faf9

Shafto, P. *Why Big Tech Companies Are Open-Sourcing Their AIS* [Internet]. IFLScience. [cited 2018 Sep 3]. Available from: https://www.iflscience.com/technology/why-big-tech-companies-are-open-sourcing-their-ai-systems/

Internet sites to visit

The FACIL website

The Déclaration des communs numériques de FACIL

4.3.

RECOMMENDATIONS FROM THE NEW GENERATION OF RESEARCHERS

This exercise was particularly fruitful, and the three briefs provide relevant recommendations on responsible AI.

The first policy brief examines the consequences of a poor evaluation of AI capabilities, a problem that is an integral part of the first problem proposed on system security and integrity. The purpose of this brief is to promote security and protection of Canadians and, more specifically with respect to embedded systems. The brief recommends creating a Canada-wide organization for certifying embedded systems that use AI (ARBIA) and developing a no-fault liability plan. The brief is notable for the relevance of its recommendations, which refer to existing mechanisms (MEI[26] and integration of the recommendations into Quebec's Digital Strategy), as well as for proposing an original mechanism for protecting citizens (either the state or the businesses that have developed such embedded systems will be liable for the damages arising from an accident).

The second brief, entitled "Fake news, real issues: educating ourselves to confront the issues" [translation], examines the problem of information on the Internet that has been manipulated using AI (in particular, the creation of fake news). Here the students underscored the need to encourage more critical thinking. Their recommendations are in line with the importance of educating Quebec citizens about media and information. Their analysis led to two main recommendations:

[24] Ruzé E. *La constitution et la gouvernance des biens communs numériques ancillaires dans les communautés de l'Internet. Le cas du wiki de la communauté open-source WordPress*. Management & Avenir. 2013;(65):189–205.

[25] Crosnier HL. *Une bonne nouvelle pour la théorie des biens communs*. Vacarme. 2011;(56):92–4.

[26] Formerly MESI (*Quebec's Ministère de l'économie*, de la science et de l'innovation), which was renamed MEI (*Ministère de l'économie et de l'innovation,* or the department of economics and innovation) on the day that the brief was written.

1. the Quebec public should be warned about fake news (vigilance), including from specialized AIS, and

2. Quebec teachers should be provided with the tools they need to raise awareness about fake news through the education system, starting in primary school.

In order to ensure that these mechanisms are implemented, they recommend creating a Quebec Digital Vigilance Committee, under the aegis of the International Observatory on the Societal Impacts of Artificial Intelligence and Digital Technologies, and integrating educational processes into Quebec's Digital Action Plan.

The third brief examines the problem of AI governance, exploring issues related to the gaps in current policies and regulations concerning the AI sector. Here the challenge is to find a form of AI governance that will best respond to the needs of the various actors affected (businesses, citizens and public institutions). The main objective of this policy brief was to *provide a method for working on and thinking about AI governance* in order to appropriately address the issues raised: the inadequacy of current policies, citizens' concerns, the lack of clarity when an incident occurs, and the absence of methods for managing AI-related problems. Several potential solutions were proposed. The students recommended creating an independent (provincial) organization to regulate the use of common data that will be responsible for, among other things, implementing educational mechanisms, as well as adjusting current laws and regulations to address the new technological realities. The brief also proposed integrating a component into the organization's mission so that it will be constantly in phase with the market. The organization in question, named "Educ'AI" in the oral presentation, would be set up as a think tank.

Summarizing, the students recommend implementing an independent organization to manage AI, with a variety of responsibilities (involving the creation of a certification or a system of vigilance), as well as educational processes. These recommendations are consistent with those formulated during other co-construction activities. While implementing an independent organization or educational mechanisms is aligned with the Declaration's recommendations for public policies, the recommendation to create a system of responsibility merits further analysis. Echoing the recommendations made at the citizen forums last winter on implementing insurance mechanisms that would set parameters for the sharing of responsibility when there is fault (see *Part 3: Summary report of the recommentations from the winter co-construction workshops*), this recommendation suggests a full-fledged analysis of general and criminal responsibility for the impacts of AI.

Lastly, the relevance of this activity has led us to support the CIÉ in its recommendation that more opportunities should be created for graduate students to be trained in non-academic professional activities and, more specifically, in political participation.

5. CONCLUSION

These three activities allowed us to explore issues related to AI development under new themes (e.g. propaganda), but also to experiment with a new approach (e.g. a simulation activity).

Independent of the activity, the participants' recommendations support the need to implement training that is tailored to everyone's needs, to update the legal and regulatory framework, and to develop new knowledge on AI developments and their impacts. These recommendations also encourage the promotion of participatory governance, with an emphasis on the importance of involving the stakeholders at different key moments in the management of AI development and in policy decision making.

By opening a discussion of new potential solutions and sectoral differences, these activities suggest that co-construction deserves to be pursued beyond the work carried out by the Montréal Declaration, and they support the relevance of a public consultation on responsible AI.

APPENDIX 1
The Paris Scenarios (in French only)

DÉMOCRATIE

Fausse nouvelle dans la campagne électorale

23 mars 2022. Ce matin, Dominique B. se rend à la réunion de crise de l'Agence sur l'intégrité de l'information (AII), mise en place dans le cadre de la Loi contre la manipulation de l'information. Le président de la République sortant, candidat à sa réélection, vient de perdre 7 points dans les sondages d'intention de vote en trois semaines et la tendance à la baisse semble se confirmer. Alors qu'il était assuré de l'emporter deux mois auparavant, il est désormais dépassé par la candidate populiste de droite qui a pris la tête de la course électorale. Le tournant se situe le 2 mars, avec la diffusion sur internet d'une vidéo montrant le président de la République discuter avec le président du Mouvement des entreprises de France, en marge de son école d'été. Le président de la République assurait qu'il comprenait la situation des entreprises qui employaient des travailleurs immigrés sans papiers, qu'il était important de maintenir des bas salaires pour garantir la vitalité des petites et moyennes entreprises, et qu'il veillerait à ce que ces entreprises ne soient pas pénalisées.

La vidéo s'était vite répandue dans les réseaux sociaux et les propos du Président avaient été relayés dans les premières heures par deux grands médias, la chaîne d'information TBT et le site lefureteur.com. Le porte-parole de l'Élysée avait immédiatement démenti les propos attribués au Président et avait fait savoir que la vidéo était un faux créé par une agence étrangère qui tentait d'interférer dans les élections françaises. La technique utilisée pour créer la vidéo avait été mise au point par l'entreprise américaine Monkeypaw Productions qui avait tiré parti des algorithmes GAN (*generative adversarial networks*), élaborés par des chercheurs de l'Université de Montréal en 2014. Contre toute attente, les images créées grâce à l'IA avaient atteint un degré de réalisme stupéfiant en moins de dix ans, si bien qu'une fausse vidéo ne pouvait plus être détectée à l'œil nu.

Ni le démenti de l'Élysée, ni le mea culpa de TBT et du Fureteur, ni encore l'interdiction de diffusion de la vidéo n'avaient eu l'effet espéré. La vidéo était encore consultable sur différents sites étrangers comme le site rassvet.io. Un député du parti populiste de droite en avait profité pour accuser le Président de faire le jeu de l'immigration clandestine et de nuire aux intérêts des Français. Le nombre de gazouillis avec le mot-clic *#Presidentclandestin* avait passé la barre des 300 000 en une semaine. À un mois du premier tour des élections présidentielles, Dominique B., directrice de l'AII, doit présenter un plan pour enrayer les effets dévastateurs de cette fausse information et rétablir les conditions d'une campagne électorale saine. Mais ce matin, le sentiment d'avoir déjà épuisé toutes les solutions l'emporte à l'AII.

ENVIRONNEMENT

La cote environnementale basée sur l'empreinte de carbone

1er février 2025. Pour la cinquième année de suite les températures battent des records de chaleur dans le monde entier. La majorité des pays ayant signé l'Accord de Paris en décembre 2015 n'ont pas tenu leurs engagements en raison des impératifs économiques de court terme, malgré les mises en garde du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). En conséquence, les villes européennes du C40, le réseau des villes engagées dans la transition écologique, ont accéléré leur coopération pour proposer à leurs habitants un système de permis carbone individuel fortement incitatif, le système ÉcoFit, connecté à leur compte bancaire et aux différentes applications d'achat en ligne : dans ces villes, le prix des biens et services est affiché en euros et en carbone, et chaque citoyen doit viser 4 tonnes d'émission de carbone par an pour l'ensemble de sa consommation. Les personnes qui atteignent cet objectif augmentent leur cote environnementale calculée par l'algorithme ÉcoFit, à partir de leurs données personnelles de consommation. Cette cote leur donne un accès gratuit à de multiples services écoresponsables en transport, éducation, formation et culture.

15 juin 2025. Au moment de passer leur commande de coquilles Saint-Jacques grâce à leur réfrigérateur FrigoMax connecté, Ive et Charles, habitants du 20e arr. à Paris, découvrent ce nouveau système de points auquel ils viennent d'adhérer : Coquilles Saint-Jacques (Provenance : Pérou) : 12 € / 22 kg éq. CO2/kg*[27]. Un message d'avertissement s'affiche : « Cet achat doit rester exceptionnel. Vous ne pourrez pas tenir votre objectif annuel si vous le reproduisez souvent. » Et l'algorithme de recommandation de FrigoMax leur propose alors des coquilles Saint-Jacques de Saint-Brieuc, fraîches, qui coûtent 22,5€ mais seulement 0,25 kg éq. CO2/kg.

15 octobre 2025. Après quelques écarts, et suite aux nombreux messages d'avertissement, Ive et Charles ont fait un effort pour consommer plus sobrement grâce aux recommandations d'ÉcoFit : régime presque végétarien, nouvelle isolation de leur logement, transport en commun et en vélo, contrat d'électricité verte, choix exclusif d'applications avec *data-centers* carbone neutre : c'est qu'au bureau, tout le monde compare maintenant sa cote environnementale !

1er décembre 2025. Grâce à leurs comportements de plus en plus vertueux, Ive et Charles ont réussi à rester juste en-dessous du plafond visé : après 6 mois, ils sont chacun à 1,95 tonne de carbone pour leur consommation globale. De plus, ils ont moins dépensé monétairement, ce qui leur procure une épargne inattendue. Le couple considère alors de réaliser son projet de séjour à Cuba pour Noël et commence à consulter les sites des agences de voyage. Un message leur parvient sur leur téléphone : « Attention à l'*effet rebond* : dépenser vos économies dans un voyage annulerait tous vos efforts ! Pensez à voyager local ! »

[27] kg éq. CO2/kg = kilogramme équivalent carbone ; exprimé ici par kg de produit importé par avion.

MONDE DU TRAVAIL

Forage des données (*data mining*) RH pour optimiser l'ambiance au travail

30 octobre 2025. Pierre-André a enfin décroché un emploi dans un bon bureau d'avocats qui traite notamment du droit de l'environnement, l'un de ses domaines de prédilection.

Après trois semaines de travail, il rencontre Marco aux ressources humaines pour une séance de mentorat personnalisée. Marco fait le point sur l'intégration de Pierre-André, sur ses attentes initiales, ses difficultés, etc. Il lui explique aussi que la firme utilise désormais AmbIA+, une IA d'analyse conversationnelle qui étudie les attitudes des salariés et aide à maintenir une ambiance de travail apaisante et productive. C'est une question d'efficacité. Ainsi, tous les courriels, appels téléphoniques et prises de parole en réunion d'équipe sont analysés pour extraire un historique des humeurs et des émotions des salariés. Ces données sont ensuite rapportées à un laboratoire de recherche en psychologie.

Pierre-André est déstabilisé et même un peu inquiet, mais Marco essaie de le rassurer :

> AmbIA+ fournit une assistance individualisée, elle conseille et entraîne, mais il n'y a pas de sanction. D'ailleurs, AmbIA+ ne mémorise que la forme des interactions, et tous les échanges que vous avez eus jusqu'à présent au bureau se sont bien passés.

Tous, sauf pour le 15 et le 16 octobre derniers. Pierre-André travaillait alors sur le dossier de la nouvelle station d'épuration des eaux usées de la ville de Lille. « Selon AmbIA+, rapporte Marco, vous avez à plusieurs reprises interrompu vos collègues en réunion pour répéter les mêmes idées, ce qui a créé de la tension chez eux. Il faudrait essayer d'exposer vos arguments en une fois, lors du tour de table, pour ne pas perdre de temps. »

Mais ce n'est pas tout :

> Apparemment, l'algorithme a aussi détecté des périodes d'inactivité sur le réseau de plusieurs heures, sans aucun échange avec vos collègues. Ce n'est pas grave en soi, mais c'est mieux de maintenir le contact avec l'équipe. Est-ce que vous vous souvenez de la raison de cette inactivité ?

Pierre-André n'est plus seulement inquiet, il est embarrassé et s'interroge sur la pertinence de ces questions :

> Oui, c'est vrai, j'aime bien travailler avec un crayon sur un rapport papier et je préfère ne pas rédiger directement sur le document collaboratif en ligne… et en effet lors de la réunion du 15, j'apportais une idée nouvelle qui ne me semblait pas bien comprise et je craignais qu'on ne l'oublie. Mais est-ce vraiment un problème ?

Compréhensif, Marco répond qu'il n'y a vraiment aucun problème : « Mais ne vous déconnectez pas de l'équipe, c'est mieux pour la performance collective. Allez, on se revoit dans deux mois. Et bonne chance pour la réunion de demain ! »

APPENDIX 2
Student policy briefs

Simulation 2018, CIÉ-FRQ

Brève politique sur l'intelligence artificielle

Le présent document est le résultat d'un exercice de simulation, dont l'objectif était d'acquérir des compétences en rédaction et en communication publique. Étant donné le contexte pédagogique dans lequel cette note a été produite, elle n'a pas la vocation, dans les faits, d'être adressée à des décideurs ou à des acteurs de la fonction publique.

La Déclaration de Montréal a choisi de publier ces brèves afin de représenter fidèlement le résultat d'un travail réalisé en 6 heures par les étudiants de la relève et montrer la pertinence d'un tel exercice.

Problématique 1 : Sécurité publique et intégrité des systèmes

Sous-problématique 4 : Les conséquences engendrées par une mauvaise évaluation des capacités de l'intelligence artificielle

Document rédigé par :

Joël Simoneau

Jérôme Gélinas Bélanger

Fidele Ndjoulou

Moumouni Ouiminga

À l'intention du Gouvernement du Canada

Titre de la brève

Pour l'établissement d'un système de responsabilité de l'intelligence artificielle dans les biens de consommation

Cette brève politique expose une démarche qui vise à promouvoir une intelligence artificielle responsable pour la sécurité et la protection des Canadiennes et Canadiens. Elle porte spécifiquement sur les systèmes embarqués. Les recommandations énoncées sont :

1. La mise sur pied d'un organisme pancanadien de certification des systèmes embarqués[1] utilisant l'intelligence artificielle
2. Le développement d'un régime de responsabilité sans faute

De manière concrète, le gouvernement devrait s'atteler dans un premier temps à la création d'un organisme fédéral responsable de la certification obligatoire des systèmes embarqués utilisant l'IA. Dans un deuxième temps, il est impératif de mettre sur place un régime propriétaire sans faute.

Une telle politique permettra d'assurer une meilleure santé et sécurité ainsi qu'une protection légale à tous les Canadiennes et les Canadiens dans leur interaction avec des objets utilisant l'IA, tout en impliquant les entreprises privées dans le processus de la saine utilisation de l'IA.

[1] On qualifie de « système embarqué » un système électronique et informatique autonome dédié à une tâche précise, souvent en temps réel, possédant une taille limitée et ayant une consommation énergétique restreinte. www.futura-sciences.com/tech/definitions/technologie-systeme-embarque-15282/

La société canadienne à l'ère du développement de l'intelligence artificielle

Notre société est en train de vivre une transformation globale basée sur l'évolution du numérique. Autant celui-ci véhicule des informations à une vitesse précédemment inimaginable, qu'il transforme notre rapport avec les objets. Les avancées technologiques récentes en intelligence artificielle (IA) permettent d'imaginer un futur imminent où certaines tâches avec prises de décision redondantes seraient attribuées à des logiciels conçus expressément pour cette fonction. Les véhicules autonomes sont déjà au coin de la rue, les dispositifs médicaux intelligents sont derrière les portes des universités. Une mauvaise médicamentation ou des accidents automobiles sont des dangers qui tendent à être réglés par l'utilisation intelligente et sécuritaire de l'IA, mais il faut aussi s'assurer qu'elle n'en devient pas la cause. Cela représente des inquiétudes énoncées par les Canadiennes et les Canadiens à travers les travaux de la Déclaration de Montréal pour un développement responsable de l'IA.

La régularisation des systèmes embarqués, soit un appareil physique contenant un logiciel utilisant une IA, devrait être un projet d'importance pour le gouvernement canadien. Ceux-ci représentent une implémentation physique et commercialisable d'un produit d'IA, et il serait important d'en assurer une règlementation en amont de leur arrivée prochaine sur le marché canadien. Une prise de décision proactive et l'installation d'un cadre réglementaire permettrait l'encadrement des IA pouvant avoir un impact physique direct sur le peuple canadien.

Ce document propose l'instauration d'un organisme réglementaire de certification des systèmes embarqués utilisant l'IA et d'un régime de responsabilité basé sur le propriétaire sans faute. L'organisme permettrait d'encadrer les normes de sécurité de conception et d'utilisation des systèmes, et le régime permettrait de définir exactement le rapport de responsabilité dans le but de protéger les Canadiennes et les Canadiens, autant légalement qu'au niveau de leur santé et bien-être. La combinaison de ces deux mesures encadrera les systèmes embarqués, de leur commercialisation jusqu'à leur utilisation, ce qui maximisera les impacts positifs du développement de l'IA, en réduisant ses effets néfastes.

Constats et pistes d'action sur le développement de l'intelligence artificielle au Canada

1. Organisme de certification

À l'heure présente, aucun cadre législatif n'existe quant à l'utilisation de l'intelligence artificielle intégrée à des systèmes embarqués au Canada. Ce flou juridique pose un certain nombre de défis pour les différents paliers de gouvernement, notamment le gouvernement fédéral, relativement à leur capacité de structurer la mise en marché et la régulation de ces objets au pays. De façon plus générale, ce manque de structure à ce niveau engendre des complexités juridiques en termes d'évaluation du risque que présentent ces technologies pour le public, mais aussi en termes de l'attribution du poids de la responsabilité advenant un incident découlant de l'utilisation d'une technologie basée sur l'IA.

Face à ces défis, il apparaît nécessaire pour l'État canadien de créer un organisme réglementaire de certification des systèmes embarqués utilisant l'IA, l'office de réglementation nommé l'Agence de Réglementation sur les Biens utilisant l'Intelligence Artificielle (ARBIA), à vocation interdisciplinaire et agissant comme pilier décisionnel. Cet organisme possédera trois principaux axes d'action afin de parvenir à structurer la réglementation de l'IA à l'échelle canadienne: 1) l'investissement dans la recherche et l'innovation permettant le développement de balises législatives basées sur des connaissances techniques, 2) l'instauration de comités experts possédant une bi-spécialisation reposant sur l'IA et leur propre champ d'expertise à l'intérieur des différents ministères pouvant être éventuellement affectés par le développement de l'IA et 3) le développement d'une plateforme réglementaire encadrant la mise en marché et le régime propriétaire sans faute.

L'implication directe du gouvernement canadien dans les cas de problématique de bien utilisant l'IA permettra d'assurer une veille scientifique et sécuritaire proactive, et de protéger légalement les consommateurs canadiens, qui n'auront pas à subir des procès-bâillons.

2. Régime de responsabilité sans faute

On entend par responsabilité l'obligation de répondre d'un dommage devant la justice et d'en assumer les conséquences notamment civiles et pénales envers la victime et/ou la société. Dans un régime de responsabilité sans faute, le gouvernement du Canada sera responsable des accidents physiques ou matériels causés par un bien matériel utilisant l'IA. Dans le cas d'un bien non conforme au processus de certification, le gouvernement canadien peut intenter des actions contre le fabricant.

L'implication directe du gouvernement canadien dans les cas de problématique de bien utilisant l'IA permettra d'assurer une veille scientifique et sécuritaire proactive, et de protéger légalement les consommateurs canadiens, qui n'auront pas à subir des procès-bâillons.

2.1 Secteurs d'activités concernés

L'intelligence artificielle s'applique à plusieurs secteurs d'activité notamment la santé, l'éducation, la sécurité, l'agriculture. Cependant, cette brève politique touche de manière spécifique l'automobile autonome, les dispositifs médicaux et la domotique.

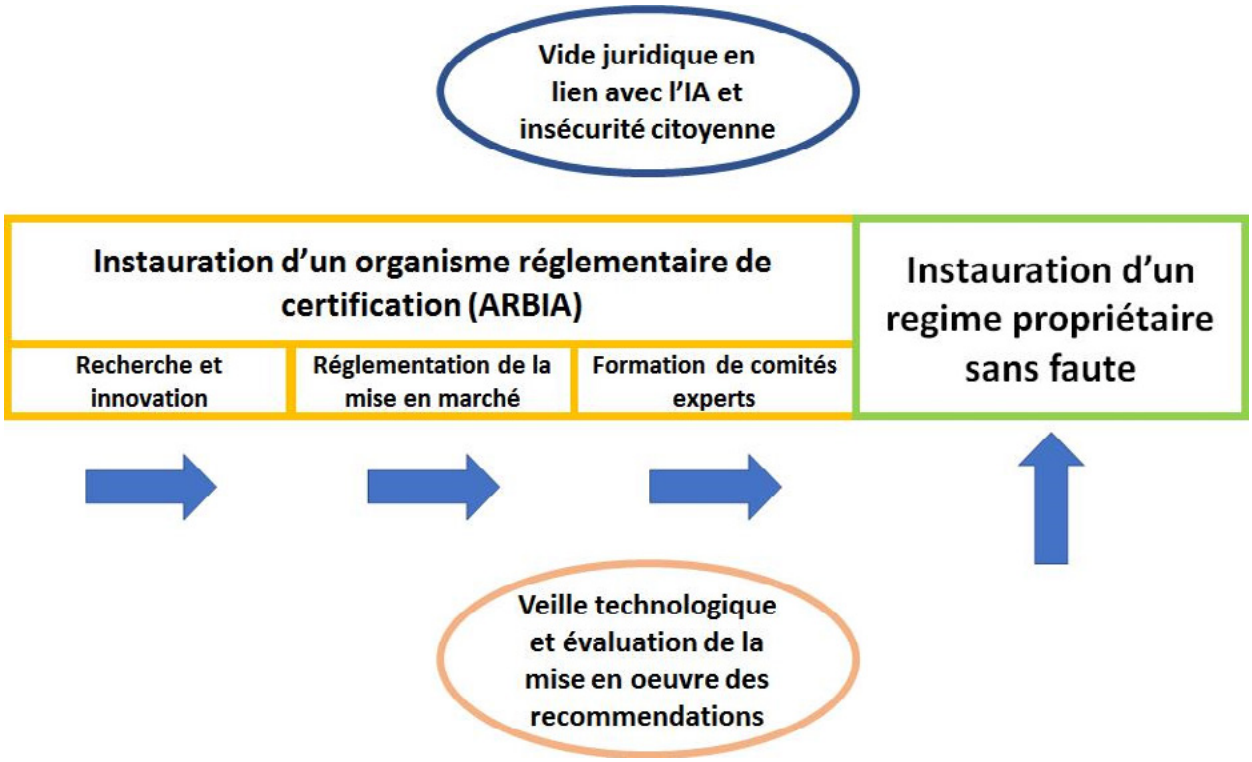
2.1.1 Automobiles autonomes

Avec le développement de l'intelligence artificielle, le secteur de l'automobile a connu une transformation radicale. Une nouvelle catégorie d'automobile dite automobile autonome est mise sur le marché. Néanmoins, ces voitures ont déjà causé des accidents aux États-Unis, par exemple l'accident mortel causé par une voiture autonome en Floride en mars 2018. Compte tenu de l'absence des règles spécifiques à la circulation des automobiles autonomes, et dans le souci d'apporter une meilleure protection aux citoyens, il est impératif de définir un cadre réglementaire au niveau fédéral. Ce cadre va fixer la responsabilité des parties prenantes, à savoir l'État et les compagnies propriétaires des voitures autonomes. Une faute liée aux défaillances est une faute de la compagnie responsable et propriétaire de la voiture, tandis qu'une utilisation faite par l'individu utilisateur est une faute de sa part.

2.1.2 Dispositifs médicaux

Pour faciliter la vie des personnes vivant avec le diabète, une pompe à insuline a été développée. C'est un système intégré composé de trois parties (boîtier,

composantes électroniques, cathéter) qui libère automatiquement de l'insuline. Le fonctionnement de ce système nécessite impérativement une formation du patient, une autosurveillance glycémique et un suivi médical rapproché. Cela engendrera une responsabilité du patient en cas de mauvaise utilisation de sa part. Les conséquences pourraient être énormes au point d'engendrer d'éventuels cas de décès.



Retombés et recommandations

Dans l'objectif d'assurer la santé et la sécurité des Canadiennes et Canadiens face aux systèmes embarqués utilisant l'IA et de maximiser les impacts positifs du développement de l'IA, nous émettons les recommandations suivantes.

Recommandation 1

Création de l'Agence de Réglementation sur les Biens utilisant l'Intelligence Artificielle (ARBIA)

Retombées :

- Coût nul pour le gouvernement canadien

  Le financement de l'organisme de certification et des actions légales sera couvert par une licence de fabrication des systèmes.
- Fiabilité des systèmes d'IA pour la santé et sécurité des Canadiennes et Canadiens.

  Par le respect de normes définies par des comités experts, normes qui seront mises à jour selon les cas vécus.

Recommandation 2

Instauration d'un régime propriétaire sans faute

Retombées :

- Accessibilité judiciaire améliorée dans le cas de faute des fabricants.

  Le système sans faute donne la responsabilité de la poursuite judiciaire au gouvernement canadien, qui a plus de ressources que les citoyennes et citoyens individuellement.
- Promotion de l'implication sociale des entreprises.

  Considérant leur responsabilité directement impliquée, les fabricants vont être encouragés à développer des mécanismes d'utilisation sécuritaire de leurs produits.
- Veille technologique et sécuritaire du gouvernement canadien

  Considérant l'implication directe du gouvernement canadien dans les processus judiciaires, celui-ci assure une veille permanente dans la gestion saine des IA.

Bibliographie :

[1] Pour aller plus loin voir Palmer, Vernon. « Trois principes de la responsabilité sans faute » (1987) 39:4 Revue internationale de droit comparé; Mémeteau, Gérard. « Un point sur la responsabilité civile du fait des prothèses » (2013) 2013:123 Médecine & Droit 175-180; Jacob, Julien. « Prévention des risques technologiques à l'aide de la responsabilité civile en présence d'une innovation à double impact » (2013) 202:1 Économie & amp; prévision 1-18.
[2] https://diabetnutrition.ch/les-traitements/la-pompe-a-insuline-quest-ce-que-cest/
3https://ici.radio-canada.ca/info/videos/media-7560667/premier-accident-mortel-impliquant-une-voiture-autonome source consultée le 18 octobre 2018

Simulation 2018 dans le cadre des J2R

« Politique et intelligence artificielle »

Le présent document est le résultat d'un exercice de simulation, dont l'objectif était d'acquérir des compétences en rédaction et en communication publique. Étant donné le contexte pédagogique dans lequel elle a été produite cette note, elle n'a pas la vocation, dans les faits, d'être adressée à des décideurs ou à des acteurs de la fonction publique.

La Déclaration de Montréal a choisi de publier ces brèves afin de représenter fidèlement le résultat d'un travail réalisé en 6 heures par les étudiants de la relève et montrer la pertinence d'un tel exercice.

Fausses nouvelles, vrais enjeux : s'éduquer pour y faire face

Présenté aux membres du jury

Par

Jean Clairemond César, étudiant au doctorat en éducation, Université de Sherbrooke
Isabelle Dufour, inf., candidate au doctorat, Université de Sherbrooke
Gaël Grissonnanche, post-doctorant en physique, Université de Sherbrooke
Philippe Lebel, doctorant en microbiologie, Université de Montréal

18 octobre 2018

Tables des matières

Tables des matières 2
But de la brève 3
Couverture (1 page) 4
Introduction (1 page) 5
Données probantes et analyse (3 pages) 6
Répercussions sur les politiques et recommandations (1 page) 9
Tableau 1. Grille d'évaluation des brèves politiques 10

But de la brève

Émettre des recommandations auprès d'un décideur public en vue d'offrir une ou plusieurs pistes de solution à un problème spécifique découlant d'une des trois problématiques décrites dans le document élaboré par l'équipe de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. Une problématique sera attribuée par équipe et les participants devront identifier les éléments suivants :

- Le **problème**
- La ou les **solution**(s) recommandée(s) et les **répercussions** sur la population visée et non visée
- Le **décideur public** impliqué
- Les **facteurs environnementaux** pouvant faire obstacle à la mise en œuvre de la ou des solution(s) recommandée(s).

Couverture (1 page)

La première page présente une synthèse de la brève politique. Elle présente la pertinence de la brève et ses grandes lignes, les conclusions clés et la marche à suivre.

Cette brève politique présente le problème des fausses nouvelles sur l'internet. Aujourd'hui la proportion de Canadiens qui consomme de l'information en ligne a dépassé celle des médias traditionnels. L'efficience de cette technologie repose sur l'intelligence artificielle (IA) en offrant des contenus filtrés selon le comportement et l'intérêt de l'utilisateur. De nos jours, plusieurs acteurs sociopolitiques ont levé le drapeau rouge sur cet enjeu de société. D'un autre côté, les grandes entreprises de médias sociaux telles que Google, Facebook, Amazon et tant d'autres proposent déjà des mesures pour limiter le potentiel propagandiste de leur algorithme, et la définition d'une fausse nouvelle ne fait pas consensus. La pertinence de notre brève se situe dans la nécessité d'augmenter l'esprit critique au sein de la population québécoise. L'absence d'esprit critique peut occasionner plusieurs problèmes en éducation et en santé et dans d'autres domaines. Destiné aux ministres concernés par la stratégie numérique, ce document présente plusieurs recommandations sur l'importance de l'éducation aux médias et à l'information au Québec.

Introduction (1 page)

Cette section décrit l'objectif principal de la brève et le problème politique. Elle établit un lien entre les données probantes et le problème.

L'avènement de l'internet apporte aux citoyens la démocratisation de l'accès à l'information à travers des moteurs de recherches intelligents et des médias sociaux. Aujourd'hui, la proportion de Canadiens consommant de l'information en ligne a dépassé celle des médias traditionnels. L'efficience de cette technologie repose sur l'intelligence artificielle (IA) en offrant des contenus filtrés selon le comportement et l'intérêt de l'utilisateur. Tandis que certaines études montrent que cette exposition partiale à l'information tend à engendrer chez l'utilisateur une confirmation systématique de sa pensée, d'autres en revanche arguent que celui-ci n'a jamais été exposé à une telle diversité de sources lorsque comparé à la presse écrite, à la télévision, à la radio, etc.

C'est dans ce contexte qu'émerge sur la scène internationale la notion de fausse nouvelle comme un enjeu de désinformation massive dans une société démocratique. L'usage d'IA comme en a fait la firme Cambridge Analytica aux États-Unis a montré au monde le niveau de déstabilisation sociétale que cette technologie peut engendrer. Alors que les grandes entreprises de médias sociaux telles que Google, Facebook, Amazon et tant d'autres proposent déjà aujourd'hui des mesures pour limiter le potentiel propagandiste de leur algorithme, la définition d'une fausse nouvelle ne fait pas consensus. En effet, selon le Global News, près de 58% des Canadiens définissent celle-ci comme une histoire pour laquelle les faits sont faux. Cependant, 46% l'emploient pour désigner les nouvelles de journaux et les discours de personnalités politiques n'exprimant qu'un unique côté des faits. Encore, ce même chiffre désigne le pourcentage pensant que ce terme est uniquement utilisé par les politiciens pour discréditer les médias qui les critiquent. À l'autre bout du spectre, des actions pour valoriser l'esprit critique de l'utilisateur demeurent une avenue qui doit être envisagée.

La problématique amenée par l'essor des fausses nouvelles dans les médias est importante et est susceptible d'avoir un impact important sur la population. À cet égard, l'objectif de cette brève est d'augmenter la sécurité de la population et leur éducation face aux fausses nouvelles.

Données probantes et analyse (3 pages)

Cette section représente le cœur de la brève politique. La qualité de cette section est jugée par la pertinence des données présentées, des interprétations tirées de ces données, ainsi que de leurs apports et de leurs limites. Elle peut contenir des graphiques, des tableaux et des schémas.

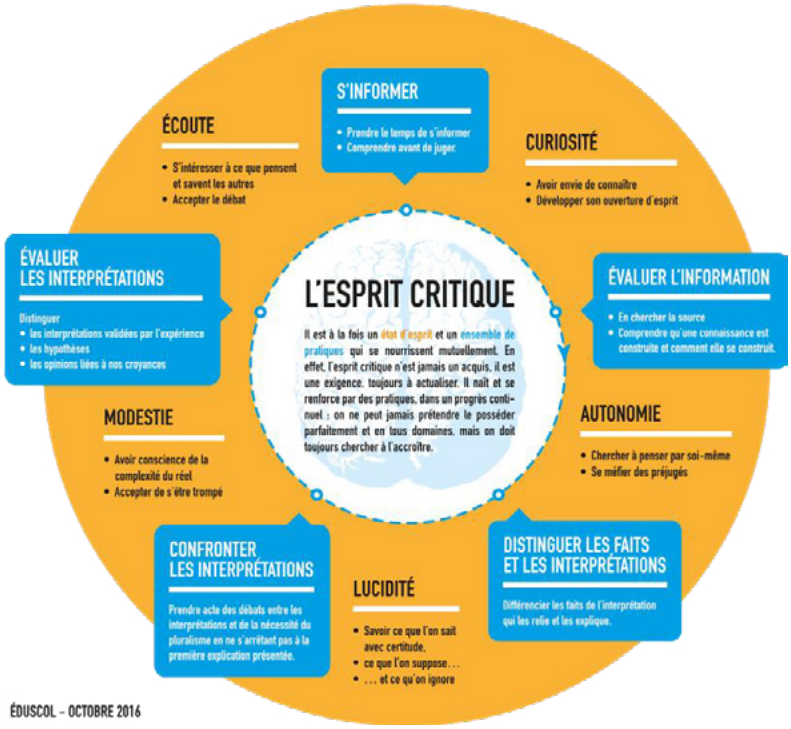
Selon Jeff Yates, expert québécois de la question, une fausse nouvelle se définit comme « une information soit carrément fausse, détournée, exagérée ou dénaturée à un point tel qu'elle n'est plus véridique, et présentée comme une vraie nouvelle dans le but de tromper les gens. Cela peut être fait pour générer des clics et des partages sur les réseaux sociaux, pour atteindre des objectifs quelconques (politiques, idéologiques, économiques, etc.) ou simplement pour se moquer de la crédulité des lecteurs ». Sujet de débats socio-économiques, les fausses nouvelles dans les médias ont connu un essor marqué durant les dernières années, et principalement avec le développement de l'IA. En effet, des méthodes associées à l'IA sont utilisées par les sites de médias sociaux et peuvent procéder de façon automatique à la diffusion de fausses nouvelles.

En Amérique du Nord, environ 60% de la population croit que la dispersion de telles informations dans les médias cause de la confusion. Les enfants et les adolescents sont particulièrement à risque d'attribuer du crédit et de participer à leur diffusion. Le manque de consensus dans la définition d'une fausse nouvelle contribue à l'incertitude vécue par la population, 45% des Canadiens en ayant une vision erronée. À l'ère numérique, la nécessité de développer une pensée critique concernant les informations transmises par les médias se positionne donc comme un enjeu central.

Selon Vallerand, la pensée critique est une pensée responsable qui s'appuie sur des critères et qui est sensible au contexte et aux autres (Vallerand, 2016). L'IA pourrait être utilisée pour développer l'esprit critique des jeunes et les former au doute constructif. Les jeunes du Québec apprendront comment mettre en perspective une information diffusée sur le web. En ce sens, les décideurs politiques donneront les moyens nécessaires pour y arriver.

En France par exemple, le développement de l'esprit critique est au centre de la mission assignée au système éducatif français, comme le présente le modèle de l'esprit critique d'Eduscol. Il est renforcé par l'attention désormais portée à l'éducation aux médias et à l'information. Le travail de formation des élèves au décryptage du réel et à la construction, progressive, d'un esprit éclairé, autonome et critique est essentiel. L'esprit critique est une compétence essentielle du citoyen et de la citoyenne du 21^{e} siècle. Analyser une source, mettre en perspective une image ou une information, en extraire l'essentiel, critiquer le contenu, se questionner sont autant de savoirs numériques nécessaires à l'exercice d'une citoyenneté avisée. Une étude faite en 2017 au Royaume-Uni a montré que seulement 4 % de la population testée avait été capable d'identifier correctement les vraies des fausses nouvelles. Ce résultat est inquiétant, notamment en termes de sécurité publique. Pensons par exemple au mouvement anti-vaccination qui

cause un retour en force de maladies mortelles, tel que la coqueluche aux États-Unis, malgré qu'il a été démontré depuis longtemps que les causes de ce mouvement sont fausses.



Selon les experts de la pensée critique, Christopher DiCarlo et l'auteur du *Petit cours d'autodéfense intellectuel*, Normand Baillargeon, la solution passe par l'éducation. En effet, il est préférable que l'école forme une jeunesse plus critique de ce qu'elle consulte plutôt que de faire confiance aux grandes entreprises privées du web pour autocensurer leur contenu.

D'ailleurs on retrouve plusieurs initiatives, ailleurs comme chez nous, qui vont dans ce sens. En France, plus précisément en Haute-Savoie, des enseignantes ont créé une habitude locale où elles prennent une heure par semaine pour sensibiliser leurs élèves à détecter les fausses nouvelles. Cette initiative, accueillie avec enthousiasme par les élèves, semble rapidement porter fruit puisque ces jeunes de 10 ans ont déjà développé les réflexes de vérifier d'où proviennent des images-chocs qui publicisent de fausses nouvelles sensationnalistes, par exemple.

Plus près de chez nous, depuis mai 2018, un nouveau programme s'implante dans les écoles ontariennes : Actufuté. Ce programme se veut une collaboration entre la Fondation pour le journalisme canadien et l'organisme CIVIX qui est responsable du programme Vote étudiant. Ce dernier prend vie autour des périodes d'élections et encourage la participation citoyenne des 9 à 19 ans. C'est dans ces périodes riches en nouvelles qu'Actufuté viendra aider les élèves à démystifier le vrai du faux.
Au Québec, « École branchée », un organisme sans but lucratif (OSBL) propose des outils aux enseignantes et aux enseignants pour intégrer ces considérations dans leur programme de tous les jours. Malheureusement, à ce jour, seulement 15 % à 20 % du corps enseignant est rejoint par l'organisme. Démonstration qu'une intervention gouvernementale est nécessaire pour offrir une protection équitable à tous nos jeunes contre ce fléau. Ce faisant, la jeunesse pourra aussi transmettre cette information et conscientiser ses proches à la problématique.

Pour y arriver, le Plan d'action numérique en éducation et en enseignement supérieur, annoncé à l'été 2018, prévoit quelque 900 millions de dollars pour, justement, préparer la génération de demain à ce nouvel environnement numérique. Le gouvernement du Québec pourrait ainsi soutenir les services d'« École branchée », voire même intégrer son contenu au cursus normal de l'éducation primaire et secondaire.

L'intégration de cette notion d'éducation directement au cursus scolaire vient contrer l'obstacle environnemental principal. Le désir des professeurs d'assurer le développement de leurs étudiants pourra aussi agir à titre de facilitateurs.

Répercussions sur les politiques et recommandations (1 page)

Cette section présente les recommandations proposées et les répercussions anticipées. Ces recommandations et ces répercussions peuvent s'organiser autour de thèmes, de parties intéressées ou d'échéancier.

L'argumentaire soulevé met en lumière plusieurs défis soulevés par l'IA. Entre autres, elle contribue à la diffusion de masse de fausses nouvelles. Cela cause de la confusion au sein de la population, une perte de confiance envers les sources d'information. Les jeunes et les adolescents sont particulièrement sensibles aux fausses nouvelles, leur capacité de raisonnement et leur esprit critique étant en construction. L'implication des instances gouvernementales est donc primordiale pour assurer la protection et l'éducation des populations, et particulièrement des jeunes, sur la problématique des fausses nouvelles. Les recommandations adressées font appel à des notions de vigilance et d'éducation.

Vigilance

Notre recommandation :

- Alerter la population québécoise sur la dissémination de fausses nouvelles

L'objectif étant de favoriser le développement et l'utilisation d'IA spécialisée pour détecter les fausses nouvelles diffusées sur les médias sociaux.
Cette initiative s'inscrit également en parallèle avec le Détecteur de rumeurs, où les alertes du CVMQ, en cas de détection d'une fausse nouvelle de grande importance, pourront être diffusées.
La création du Comité de vigilance numérique du Québec sera annexée à l'Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique.

Éducation

Notre recommandation :

- Offrir des outils au corps enseignant québécois pour intégrer la conscientisation face aux fausses nouvelles dans l'éducation, dès l'école primaire.

Cette mesure aura deux buts : préparer directement cette génération à affronter le fléau des fausses nouvelles et les inciter à répandre ces bonnes pratiques auprès de leurs proches.

Grâce au financement déjà prévu pour le Plan d'action numérique en éducation et en enseignement supérieur ainsi qu'aux initiatives déjà en place, il sera possible de protéger la population québécoise sans investissement supplémentaire et sans réinventer la roue. À court terme, la promotion de ces outils auprès des enseignantes et des enseignants aura déjà un impact et il sera possible de penser intégrer ces enseignements au cursus normal à moyen terme.

Tableau 1. Grille d'évaluation des brèves politiques

Critères	Tous les points	- 1 p	- 2 p	- 3 p
Couverture	La synthèse présente la pertinence de la brève et ses grandes lignes, les conclusions clefs et la marche à suivre.	Des éléments sont manquants	La synthèse est manquante	
Introduction	Cette section décrit très bien l'objectif principal de la brève et le problème politique. Elle établit un lien entre les données probantes et le problème.	Cette section décrit bien l'objectif principal de la brève et le problème politique. Elle établit un lien entre les données probantes et le problème.	Cette section décrit convenablement l'objectif principal de la brève et le problème politique. Elle établit un lien entre les données probantes et le problème.	Cette section est absente
Données probantes et analyse	Cette section est très pertinente au regard du problème; Les interprétations sont justes et convaincantes; Les facteurs environnementaux (socio-politico-économico-culturels) sont très bien pris en compte dans la possible intégration des recommandations; Les apports et les limites sont très bien identifiés.	Cette section est pertinente au regard du problème; Les interprétations sont justes; Les facteurs environnementaux (socio-politico-économico-culturels) sont bien pris en compte dans la possible intégration des recommandations; Les apports et les limites sont bien identifiés.	Cette section est plutôt pertinente au regard du problème; Les interprétations sont plutôt justes; Les facteurs environnementaux (socio-politico-économico-culturels) sont pris en compte dans la possible intégration des recommandations; Les apports et les limites sont identifiés.	Cette section n'est pas pertinente au regard du problème; Les interprétations sont erronées; Les facteurs environnementaux (socio-politico-économico-culturels) ne sont pas pris en compte dans la possible intégration des recommandations; Les apports et les limites ne sont pas correctement identifiés.
Répercussions et recommandations	Les recommandations proposées sont très pertinentes et les répercussions anticipées très bien identifiées.	Les recommandations proposées sont pertinentes et les répercussions anticipées sont bien identifiées.	Les recommandations proposées sont plus ou moins pertinentes et les répercussions anticipées plus ou moins bien identifiées.	Les recommandations proposées ne sont pas pertinentes et les répercussions anticipées ne sont pas bien identifiées.
Qualité de la présentation orale	La présentation de la brève est très convaincante.	La présentation de la brève est convaincante.	La présentation de la brève est peu convaincante.	La présentation de la brève n'est pas convaincante.
Total des points	/15			

Simulation 2018 dans le cadre des J2R

« Politique et intelligence artificielle »

Le présent document est le résultat d'un exercice de simulation, dont l'objectif était d'acquérir des compétences en rédaction et en communication publique. Étant donné le contexte pédagogique dans lequel a été produite cette note, elle n'a pas la vocation, dans les faits, d'être adressée à des décideurs ou à des acteurs de la fonction publique.

La Déclaration de Montréal a choisi de publier ces brèves afin de représenter fidèlement le résultat d'un travail réalisé en 6 heures par les étudiants de la relève et montrer la pertinence d'un tel exercice.

Gouvernance publique, privée ou participative : les communs numériques

Présenté aux membres du jury

Par

Thomas Bousquet
Alexandre Côté, PhD(c)
Christian Kouakou, PhD(c)

18 octobre 2018

Tables des matières

Simulation 2018 dans le cadre des J2R 1
« Politique et intelligence artificielle » 1
Tables des matières 2
Couverture 3
Introduction 4
Données probantes et analyse 5
Le problème de l'immigration discriminante 5
La confidentialité des données 5
La stratégie du Québec en matière d'intelligence artificielle 6
Répercussions sur les politiques et recommandations 7

Couverture

Vue d'ensemble

L'évolution rapide de la technologie et la science entourant l'intelligence artificielle (IA) exposent certaines brèches dans les politiques et les règlementations actuelles qui concernent ce secteur de développement. Le gouvernement doit se pencher sur ce problème qui soulève de vives inquiétudes pour la population qui s'interroge sur la protection de sa vie privée. L'inquiétude reste présente du côté des entreprises privées qui elles, ont incessamment besoin d'alimenter leur système d'IA avec des données de plus en plus complexes et précises. Le défi du gouvernement est de trouver une forme de gouvernance de l'IA qui répond au mieux aux besoins des différents acteurs concernés.

L'objectif principal de cette brève politique proposée par notre Groupe de travail ponctuel sur l'utilisation de l'IA et des données communes est donc de *fournir une méthode de travail et de réflexion* afin de répondre adéquatement à ces interrogations, en s'assurant de considérer les intérêts distincts de la population et du secteur privé.

Intérêts des acteurs impliqués

Population	- Protéger ses données personnelles - Être rassuré par l'indépendance des instances faisant usage de ses données
Gouvernement	- Assurer la protection du public - Stimuler la croissance économique du secteur technologique
Privé	- Connaître une croissance économique stable - Développer et améliorer les connaissances concernant l'IA

Problèmes soulevés par ces intérêts distincts :

- Politique actuelle mal adaptée
- Inquiétudes des citoyens
- Flou dans les responsabilités lors d'incidents
- Absence de méthodes pour gérer les problèmes liés à l'IA

Plusieurs pistes de solutions

- Création d'un organisme de régulation indépendant
- Favorisation d'une responsabilité partagée relativement aux données communes
- Ajustement des lois et réglementations en vigueur afin de les adapter aux nouvelles réalités technologiques

Introduction

L'évolution rapide de la technologie et la science entourant l'intelligence artificielle (IA) exposent certaines brèches dans les politiques et les règlementations actuelles qui concernent ce secteur de développement. Bien que certaines lois soient déjà en place, la vitesse de l'appareil public peut difficilement rattraper celle de la croissance technologique, et les règles en place deviennent rapidement inadaptées.

Les inquiétudes de la population

Cette inadéquation des politiques publiques en matière d'encadrement de l'IA et de l'utilisation des données servant à sa croissance inquiète la population québécoise. Une consultation récente initiée par un groupe d'experts composé, entre autres, de gens de l'Université de Montréal, de l'Université McGill et de l'Institut de valorisation des données (IVADO), a permis d'identifier certaines préoccupations clés des citoyens vis-à-vis les enjeux actuels concernant notamment :

- La responsabilité face aux données et à l'IA ;
- La protection de la vie privée des individus ;
- La valeur marchande des données partagées ;
- Les risques de mise en place d'un monopole, et de conflits d'intérêts entre les différents acteurs touchés ;
- Ainsi que l'indépendance des différents acteurs qui interviennent dans le domaine.

Les considérations face au secteur privé

La croissance économique québécoise étant de plus en plus liée aux nouvelles technologies, à l'exploitation des données et au développement de l'IA, il est important pour le gouvernement – malgré les inquiétudes soulevées – de ne pas laisser le secteur privé au dépourvu. L'accès aux données de la population est le carburant de cet important moteur économique qui doit manifestement être régulé, mais pour qui une marge de manœuvre doit être maintenue.

Le défi de la gouvernance

Les trois acteurs généraux qui sont touchés par la problématique – le gouvernement, la population et le secteur privé – ressentent déjà les impacts du manque d'ajustement des politiques actuelles. L'exemple récent des piratages de données des grands services technologiques comme Facebook et Google – utilisés par des centaines de milliers de Québécois – expose bien ce phénomène : la population est ultimement la victime, blâmant à la fois le secteur privé et le gouvernement pour les dommages encourus.

Dans cette optique, il est donc primordial pour le député délégué à la transformation numérique gouvernementale de se positionner et de répondre aux questions et inquiétudes soulevées par la société civile et le secteur privé :

1. Qui est responsable face aux données communes ?
2. Quel appareil assure la protection du public ? Fonctionne-t-il adéquatement ?
3. Quel niveau de transparence est optimal ?
4. Comment peut-on stimuler la croissance économique dans le secteur des nouvelles technologies, tout en maintenant la confiance de la population face au processus ?

L'objectif principal de cette brève politique proposée par notre Groupe de travail ponctuel sur l'utilisation de l'IA et des données communes est donc de *fournir une méthode de travail et de réflexion* afin de répondre adéquatement à ces interrogations, en s'assurant de considérer les intérêts distincts de la population et du secteur privé.

Données probantes et analyse

Le problème de l'immigration discriminante

Le gouvernement canadien serait en expérimentation de l'utilisation de l'IA pour le tri des demandes de visa et d'immigration. Cette information a été publiée dans un rapport du Citizen Lab et relayée dans les médias canadiens.

L'une des auteurs du rapport a souligné que « sans garanties et mécanismes de surveillance appropriés, utiliser l'IA pour déterminer l'immigration et le statut de réfugié est très risqué ». Il est clair que l'utilisation de l'IA pourrait aider grandement à accélérer le tri et traitement de données d'immigration. Cela ne saurait cependant et en aucun cas outrepasser la décision discrétionnaire liée au droit à l'immigration qui ne saurait être laissé à une machine ou un algorithme ; algorithme qui plus est, n'est pas sans biais car fortement dépendant des considérations des personnes qui programment. Un arbitrage est à faire entre « rapidité dans le traitement des demandes » et « sélection discrétionnaire des dossiers ».

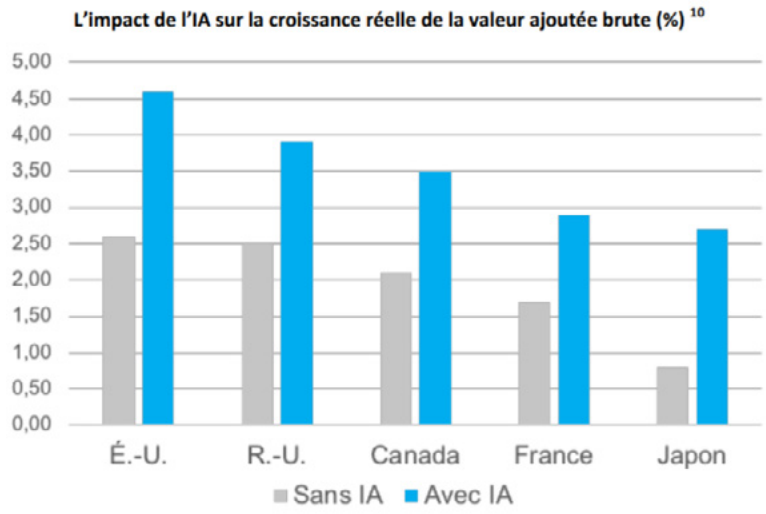
En outre, le rapport à l'immigration de l'équipe responsable de l'algorithme pourrait fortement déteindre sur le résultat final de sélection. D'une part, l'on pourrait avoir un processus de sélection moins rigoureux, voire laxiste, qui laisserait entrer sur le territoire québécois des personnes ne remplissant pas les conditions requises pour l'immigration. D'autre part, un processus plus strict pourrait ôter la possibilité aux personnes remplissant les conditions d'y accéder. Car rappelons-le, la sélection initiale aura été faite non pas par choix discrétionnaire, mais par une machine intelligente.

La confidentialité des données

Les plus gros acteurs en matière de la gestion des données d'utilisateurs, dont font partie Google et Facebook, mettent en place des actions correctives dans leur gestion des données personnelles : ces données ont une valeur monétaire importante et ne doivent être utilisées que dans le cadre où elles ont été fournies par l'utilisateur du service. Les piratages sont récurrents comme l'ont montré les évènements récents (500 000 comptes Google et 29 millions d'utilisateurs Facebook piratés) et la population est donc inquiète de l'utilisation qui peut être faite de ses données.

Les *Big Data* ont également un intérêt dans la santé des populations : des intelligences artificielles sont en développement pour aider au diagnostic médical et nécessitent donc des données sensibles, très personnelles et donc qui font partie d'un haut niveau de confidentialité. La société québécoise a donc évoqué ses inquiétudes quant à la gestion de ces données lors de l'étude qui a été menée sur une partie de la population lors de la rédaction de la Déclaration de Montréal pour l'IA responsable.

Les données personnelles sont donc classifiées selon leur confidentialité et sont donc disponibles à plusieurs niveaux. L'utilisateur sait à qui et dans quel but chaque organisme collecte des données à son sujet.
Il devient donc important qu'un comité multisectoriel neutre se charge de veiller à ce que les possesseurs de ces grandes quantités de données les utilisent de manière éthique pour éviter que des centaines de milliers d'utilisateurs québécois de ces services voient leurs données diffusées sur le web.



De nombreux pays font du développement de l'IA une priorité majeure en investissant dans des organismes de recherche visant à progresser dans ce domaine, comme le montre ce graphique issu de la communication de « Économie, science et innovation Québec » à propos de l'essor de l'écosystème québécois en intelligence artificielle publié en mai 2018.

Le Québec ressent également ce besoin et prévoit des fonds à la recherche dans ce domaine, c'est pourquoi cette volonté doit alors s'étendre à la protection des populations et de leurs données sans pour autant empêcher les acteurs privés et universitaires de progresser dans leurs recherches et leurs innovations.
D'après le projet de mettre le numérique au service du bien commun au Québec (disponible sur le site économie.gouv.qc.ca) d'ici 5 ans, le Québec prévoit mettre à la disposition de la population une transformation numérique des municipalités qui se traduit par une collecte de données continue.

De plus, le Québec prévoit également que les citoyens pourront interagir de façon numérique avec les services de santé et sociaux d'ici les prochaines années. Toutes les données nécessaires à ces services nécessitent des données sensibles sur les citoyens et il est donc très important pour la province de se mobiliser pour protéger les utilisateurs contre une mauvaise utilisation de ces données dans tous les domaines confondus, que ce soit les services publics ou bien les entreprises privées de services.

Le développement du numérique, prisé par le Québec, doit donc faire évoluer les règlements relatifs à la protection des données des citoyens en faisant travailler les entreprises leaders de l'IA conjointement avec les pouvoirs publics et les intérêts des utilisateurs.

Répercussions sur les politiques et recommandations

Nous constatons au final un manque d'adaptation des politiques actuelles face à l'IA et la gestion des données qui alimentent sa croissance. Cette problématique engendre un flou qui touche non seulement le gouvernement et la classe politique, mais également la population et le secteur privé.

Il existe actuellement un manque de clarté quant à la responsabilité de ces trois acteurs face aux données communes (1). Non seulement la *Loi sur la protection des renseignements personnels et des documents électroniques* (LPRPDE ; fédérale) et la *Loi sur la protection des renseignements personnels dans le secteur privé* (provinciale) ne répondent pas à ce manque, elles ne se sont pas non plus adaptées assez rapidement aux nouvelles réalités technologiques, ce qui risque de créer un doute dans la population quant à sa protection et la protection de ses données personnelles (2). Un certain degré de transparence (3) est évidemment requis afin de pallier cet aspect de la problématique, tout en gardant à l'esprit l'importance de ne pas entraver le développement économique du secteur technologique au Québec.

Recommandations

Dans cette optique :

1. Nous emboîtons le pas du *Citizen Lab* de l'Université de Toronto et recommandons la création d'un organisme indépendant de régulation de l'utilisation des données communes. À la différence des chercheurs torontois, nous recommandons cependant que cet organisme soit de juridiction provinciale afin de prendre en considération les particularités de la population québécoise.

2. Nous favorisons une responsabilité partagée des données communes, alimentée par ce nouvel organisme. Ce dernier serait chargé, entre autres, de l'éducation de la population en matière de protection des données personnelles et de la surveillance du secteur privé quant à l'utilisation de ces données.

3. La création de cet organisme viendrait également répondre à la deuxième question soulevée par notre analyse de la situation, soit l'identité de l'appareil veillant à la protection du public. Il serait maintenant clair aux yeux de la population qu'une entité veille à ses intérêts, entre autres en s'assurant – à travers des recommandations émises à l'endroit du gouvernement – de l'adéquation des lois et réglementations concernant l'IA et la gestion des données.

4. Nous suggérons fortement que le nouvel organisme s'assure qu'un certain niveau de transparence soit respecté, autant par le gouvernement que par le secteur privé. Le type de données utilisées, leurs sources, les buts de leur utilisation et leur portée devraient être de nature publique.

5. Nous recommandons qu'un volet économique soit intégré à la mission de l'organisme afin que celui-ci soit constamment en phase avec le marché, s'assurant que les politiques mises en place permettent d'atteindre le parfait équilibre entre croissance et respect du milieu.

6. Finalement, nous recommandons que le nouvel organisme développe une méthode de réflexion permettant d'adapter les lois et réglementations concernant l'IA et les données communes aux changements rapides et fréquents inhérents au secteur des hautes technologies.

< >
Montréal Declaration
Responsible AI_
</ >

PART 5

SUMMARY REPORT OF ONLINE SURVEYS AND PROPOSALS RECEIVED FOR THE MONTRÉAL RESPONSIBLE AI DECLARATION

TABLE OF CONTENTS

1.	**INTRODUCTION**	**214**
2.	**ONLINE SURVEY**	**215**
	Well-being (environment, caution)	215
	Autonomy	218
	Justice (equity, solidarity, diversity)	223
	Privacy (intimacy)	227
	Knowledge (publicity, caution)	231
	Democracy (publicity, diversity)	234
	Responsibility (caution)	236
3.	**SUMMARY OF SUBMISSIONS RECEIVED**	**240**
	Privacy	241
	Justice	243
	Responsibility	244
	Well-being	245
	Autonomy	246
	Knowledge	247
	Democracy	247

WRITTEN BY

MARTIN GIBERT, Ethics Counsellor at IVADO and researcher in Centre de recherche en éthique (CRÉ)

1. INTRODUCTION

In November 2017, the first step of the *Montréal Declaration for a Responsible Development of Artificial Intelligence* was launched following a convention organized by the Université de Montréal at the Palais des congrès in Montréal. The preliminary version of this Declaration, articulated around seven principles, would serve as the basis for a co-construction phase from which a new version would be created. Although the discussion workshops helped reach citizens and experts, there were also other ways to join the collective reflection: 1) by filling out an online survey accessible through the Declaration's website (www.declarationmontreal-iaresponsable.com), and 2) by sending in a proposal on one or more aspects of the Declaration. This report presents a summary of the proposals received and the answers to the survey. The report on the co-construction workshops is also available on the Declaration's website.

2. ONLINE SURVEY

The online survey consisted of 35 questions, five for each principle. A total of 83 people answered the survey, 17 of whom were anglophones. As the summary reveals, many had advanced knowledge of AI and the ethical and social issues raised by its development.

Questions are presented per order in the questionnaire, which was based on the preliminary Declaration plan. Since the revised Declaration is more complete (it is made up of 10 principles), the relevant new principles were added to those from the preliminary version.

WELL-BEING (ENVIRONMENT, CAUTION)

1. HOW CAN AI CONTRIBUTE TO WELL-BEING?

This was a general question that sparked many answers, and many varied ones at that. One recurring hope was for healthcare and assistance for the elderly or disabled. AI also seems to hold promise for reducing environmental impacts, though it was noted that "AI development has an environmental footprint (and thus a direct impact on well-being) that is often neglected, even though it is significant." Many pointed out that AI could replace humans for dangerous tasks. The aspect of "decision-making assistance" especially in the form of a personal assistant that could also assist in information searches, was also mentioned numerous times.

We expect AI to improve productivity and free us from repetitive and routine tasks as well. IT could also anticipate our needs and expectations, or simply do the vacuuming for us. One important provision: AI will improve our well-being "as long as we live in a true democracy, where it serves everyone, not only a privileged few".

SELECTED EXCERPTS

"The AI or any technology will create a lot more value for the rural population than the urban population. A single smartphone can provide immense value, and anything that can collect data is a breeding ground for AI: Better education, better farming technology (e.g. crop analysis, robot farming)."

2. CAN AN AUTONOMOUS WEAPON BE USED TO KILL A HUMAN BEING? AN ANIMAL?

An overwhelming majority of people answered "no" to this question, often very emphatically and with numerous exclamation marks. Reasons included that "killing must remain in the hands of humans, who must be fully aware of their actions". The idea of legally banning autonomous weapon systems was also mentioned many times. One survey respondent also pointed out the risk of an arms race and possible programming errors. Some respondents made a distinction: "no' for humans, "yes" for animals ("for population control"). There seems to be exceptions in some cases: a machine killing a death row inmate or a "tiger that breaks free from its cage and threatens the general population". In each case, it appears that AI should only be a tool used for killing and that in the end humans should be held responsible. One respondent, however, offered a more critical point of view and raised a valid question: if such a weapon can make a better decision than a human being, why not?"

We also noticed that this question depends largely on context: "It would be acceptable for an autonomous weapon to kill a human being or an animal in any circumstances where it would be acceptable for a human or other creature to kill a human or animal."

SELECTED EXCERPTS

"Autonomous weapons shouldn't exist, they should be banned just like chemical weapons. Humans should always be in control of a weapon; they would therefore be morally responsible for their actions."

"No! (...) a horrific scenario could ensue from an unethical manufacturer or rogue programmer who perhaps, unbeknownst to either the weapons' company or weapon purchaser, may secretly design, code & program autonomous weapons which reflect their secret views & biases as a neo-Nazi or K.K.K. supporters, for example."

"Why would you think a HUMAN should be able to kill somebody? If you have a reason, then why doesn't it apply to AI? There's no reason humans should always occupy a privileged position with respect to killing other humans. Obviously 'AI' at the moment is not even ready for consideration for this, but that's unlikely to be permanent (assuming you think anything or anybody should be killing whoever or whatever you're thinking about killing). The interesting question may be how you'll know when it's changed, and how you manage the transition."

3. SHOULD AI BE USED TO CONTROL A SLAUGHTERHOUSE?

As in the previous question, the vast majority answered "no" (though some were more in favour). The argument around normalizing violence through psychological distance, which was evident in the previous question, comes up once again: "This would distance man even more from the action of killing the animal," or even: "we must not offer humans a new level of cowardice to hide behind by delegating a morally reprehensible task to a robot". The environmental argument was also cited: "This is not the direction to take for the future of the planet and humans who live on it. I would rather see AI control greenhouses and zero CO2 emission and zero-waste buildings."

There were a few arguments in favour of such a project: avoid mistreating and creating stress for the animals, as well as improving hygiene. (One respondent nonetheless explains that a human should always oversee slaughtering operations precisely to avoid cruelty...) Some positions were mixed: AI could control the cutting and wrapping, but not the killing. Many wonder, however, if slaughterhouses are acceptable, even without AI.

SELECTED EXCERPTS

"I understand it would relieve those who are responsible for these morbid tasks. However it would be absolutely unethical."

"Slaughtering conditions may be slightly improved, but the practice itself would last longer, because it would be easier to look away."

"Yes, to the extent that the AI follows humane (and human) protocol."

"Interesting question. One side of me says 'yes', the other 'no'. What we are doing to other animals that we raise for food already has some serious ethical issues. When I read about the life of the average chicken raised for food, I was shocked. Totally automating the process of raising food, including having AI do the killing would just put the fate of these animals even more out of sight, out of mind. So, on balance, I think I am against an AI-controlled abattoir."

"Slaughterhouses already exist and won't stop existing anytime soon. AI can make sure that the method of slaughter is ethical and is done in the most humane way possible. This can also strictly ensure and maintain safety standards."

4. SHOULD WE ENTRUST AI WITH MANAGING A LAKE, A FOREST OR THE EARTH'S ATMOSPHERE?

This question elicits significant skepticism of AI, but also hope for solutions to the environmental crisis. Once more, the main idea that emerges is that the AI taking care of the environment should be "configured by responsible human beings who take conservation to heart". Some are even hopeful, especially for the climate, but from the perspective of a human/AI collaboration, rather than delegating the problem to AI. Many respondents hope for AI that cannot be corrupted or seek profit at all costs.

The risk of a malicious hijacking of AI entrusted with such a mission was also raised. There was a hint of cynicism too: "humans have destroyed nearly every natural environment they have come into contact with, so it can't really get much worse..." The theme of replacing humans also comes up: "We could have AI do everything, but we need to ask ourselves if we want humans to be assisted with reaching their full potential." There is also a democratic principle: no human or artificial entity alone should be able to decide how the environment is managed—this should be based on cooperation between all humans.

SELECTED EXCERPTS

"No, because we are sorely lacking the knowledge to be able to judge the long-term repercussions of actions taken by AI."

"It depends on what the instructions given to the IA are and how much absolute control it holds. I think AI trained on environmental systems and with ability to monitor and consider big (environmental) data could make much better decisions than any group of individuals, effectively helping to protect the environment and regenerate those that may have been affected by industry, etc."

"It could be done with the assistance of AI; but AI itself doesn't know what is good for the lake, the forest or the atmosphere. It is more efficient from the point of view of instrument rationality, but cannot determine its own goals by itself."

"At present, AI would be most useful in the collection and analysis of data."

"Eventually, machines may be more competent than people to make almost all decisions. But, if we give the machine control and stop monitoring what and how it does what it does, the ability of human beings to manage our affairs will pass out of the living memory of humans, and we will be entirely dependent upon machines. This does not seem to me to be a good future for human beings."

5. SHOULD WE DEVELOP AI CAPABLE OF EXPERIENCING WELL-BEING?

Participants hesitated on this question and answers were contradictory. Is it even possible, from a technical standpoint? Some mentioned that sentience could allow humans to control or punish AI. Others see the interest in AI being able to better understand humans (and other sentient beings) and empathize with them. Emotional intelligence also appears to be a requirement to make good moral judgments. However, simulated empathy appears sufficient, because some foresee danger in sentient AI looking after their own well-being over the functions it was assigned by humans. "AI must remain a tool that serves humans, not a quasi-human." And many wonder: "what's the point?" One respondent worries for AI: "I'd rather AI understand well-being than experience it, especially because in the notion of well-being, there's also the notion of being unwell."

SELECTED EXCERPTS

"I think it makes most sense to approach the development of general AI as the development of a calculator/tool. Developing a personified, sentient AI may be bringing new life to the world. I'm sure it would be treated fairly or with rights."

"This is a very complex question. If it experiences a sense of well-being, it will feel the need to maximize it. This is useful when rewarding learning, but will it balance its machine well-being with that of humans and other living creatures?"

"Yes, but only proportional to it accomplishing the tasks it was assigned. You could develop AI which, thanks to the satisfaction of a job well done, constantly seeks to improve, but only in the specific field in which they operate."

AUTONOMY

1. HOW CAN AI CONTRIBUTE TO HUMAN AUTONOMY?

AI presents us with an ambiguous relationship to autonomy: it makes us rely on it ("we could no longer dissociate from AI), while freeing humans from certain alienating cognitive tasks (e.g. driving a car, administrative functions), and even the need to work. What mostly shines through in the comments, however, is the positive and liberating aspect to it. The partnership model is an option, as is having AI as

a simple assistant. And the sky seems to be the limit: AI could improve the human condition, especially for people suffering from a disability, and could lead to less invasive medical care, which would also help the elderly who are losing their autonomy.

SELECTED EXCERPTS

"AI should be used to restore autonomy (physical or mental) to people with disabilities. Only a person who can entirely control the configuration of the algorithms for an AI system could gain autonomy, everyone else would lose some because they rely on decisions made by someone else."

"No system will ever help (or currently helps) human autonomy if it comes from a private company. Regulations and involvement from the public sector are essential to maintain balance."

"By freeing them from the tasks they don't want to do, and by improving their emotional state and their understanding of the world."

"HUMANs will deliberately develop AI to force other humans to follow their values or act according to their interests. And those humans will see themselves as benevolent in doing that which is the really scary part."

"AI can automate most of the trivial things that we spend a lot of time doing. Almost everything that we do without actively thinking about it can in a way be simplified or made more convenient using AI. But this also has to ensure that humans don't become too dependent on the technology, which would then handicap their life instead of providing more autonomy."

2. SHOULD WE FIGHT THE ATTENTION-CAPTURING PHENOMENON THAT COMES WITH AI BREAKTHROUGHS?

Participants were highly skeptical of this phenomenon ("I need more information"). But many highlight the risk of "technological hypnosis", "especially among teenagers". One respondent sums it up this way: "we must not become slaves to our technologies". Another suggests treating AI with AI: "we'd have to know why our attention is being captured. Income-generating goals, which drive applications such as Facebook should be blockable through other AI applications, countermeasures of sorts made available to users to fight against intrusions."

Evidently, drawing people's attention to ethical problems seems to be a good idea: "it's another way of making sure these conversations take place". And one comment that makes a lot of sense: "Yes, businesses should be prevented from manipulating people's attention in ways that people don't control or understand. That's not intrinsically related to AI; it's just that AI is a convenient, powerful and, therefore, dangerous tool for it."

SELECTED EXCERPTS

"Yes. First by educating people, then by passing legislation to impose an operational framework that reflects humanist values (truth, justice, kindness, respect, etc.)"

"All technologies, from radio frequencies, nuclear energy to

cryptography must live within a regulatory framework. Attention-seeking AI could be classified as addictive entertainment, like gambling."

"One way to combat this is awareness about the problem, the fact that this is happening is not known to many (hypothesis). And give the user proper tools to combat this: nudge the user to actually learn the skill using small dopamine hits until the user doesn't need it anymore."

3. SHOULD WE BE WORRIED IF HUMANS PREFER THE COMPANY OF AI TO THE COMPANY OF OTHER HUMANS OR ANIMALS?

No clear trend emerges from the answers. Certainly, there is concern that technology separates or isolates humans: "human beings must remain social beings" and we must ensure that humans do not forget social skills such as empathy. But from this point of view, AI "would be no worse than video games". Psychological studies are likely required to evaluate the risks of a new type of addiction. But there is also a potential benefit for people who are alone, or for certain psychological profiles: autistic children, for example, may find it easier to communicate with AI than with a human. It should remain a marginal thing, however, because "if a human no longer wants any contact with other humans, then humanity disappears". But what about paternalism? If it causes no harm to others, why prevent strong relationships between humans and AI? After all, we generally accept that certain people prefer the company of animals to that of humans.

SELECTED EXCERPTS

"If AI agents have no awareness or feelings, they can't be good company. Less so than animals even. For the moment, they are things. Machines."

"If a person has no other option, it can be a good thing. Otherwise, we're going to start having a hard time living as a community."

"No, many human beings are already trapped in relationships with objects or fictional characters (television, soap operas, social network friends). AI companionship would at least have the advantage of presenting a certain degree of interaction that could prove especially beneficial to people who are elderly or alone."

"This is a legitimate concern. It can be compared to the preference for texting as a substitute for direct human-to-human interaction."

"If you care about people's autonomy, then LET THEM MAKE THEIR OWN DECISIONS. It doesn't matter whether you're 'worried', because it's purely none of your business, full stop."

"Even if technologies like VR [virtual reality] are developed to an almost realistic level, it would only increase social isolation, and it would be detrimental in the long

run. Social security, i.e. the fact that there are people to support you and will be there with you in your time of need, is invaluable!"

4. CAN WE GIVE OUR INFORMED CONSENT WHEN FACED WITH INCREASINGLY COMPLEX AUTONOMOUS TECHNOLOGIES?

Many respondents felt this would be difficult for two reasons: the complexity of machines and the complexity of legal clauses. No one reads the terms of consent for apps or platforms when they are too complex (legalese): when people "accept" (do they really have a choice?), this consent cannot be considered truly informed. "How often do we sign off on online agreements saying we read them when we didn't?"

Systems that are secure and that inspire trust therefore still need to be created. The lack of digital literacy was also highlighted, as was the need to remedy the situation through education. That also highlights the "importance of establishing a code of ethics on which AI is built". One solution could come from AI itself: it should be able to answer our questions to help us make informed decisions. But another danger lurks: "Information presented to humans will naturally inform (and bias) decision-making. Humans are quick to assume that algorithms or information provided by statistical analysis is somehow void of bias."

SELECTED EXCERPTS

"It would be a good idea to provide a legal framework for this notion when it comes to companies and public organizations doing business in Quebec."

"It's impossible. The only thing to do is establish or restore trust with those who build or own these technologies through social or political control that satisfies the greatest number of users, by reducing the harmful use and misappropriation that the owners and designers of these technologies could be tempted to carry out."

"Probably not. I think it's already impossible to provide informed consent for digital technologies that aren't even based on AI. For example, how can we be sure the software we buy isn't spying on us?"

"For decently complex systems, the user has to be fully made aware of how the data being generated can and might be used, along with theoretical guarantees or open code base proving their claims. But for very complex systems, here, even the creator wouldn't know how the data might be used completely. But, even in the worst of the cases, rigorous proof of claims and possible benefits, analysis on a test group can help earn the trust of the user and allow the person to give consent."

"As technology advances, the demands for our consent will increase exponentially. Under those conditions, the unaided human will not be able to give truly informed consent in many of the cases where it is demanded. The proof is that we

already have become conditioned to signing off on agreements that we have not actually read or understood. The demands are only going to increase. The solution, if there is one, would involve 'loyal' AI agents assisting us."

5. SHOULD WE LIMIT THE AUTONOMY OF SMART COMPUTER SYSTEMS? SHOULD A HUMAN BEING ALWAYS HAVE THE FINAL SAY?

Many responses were positive. Human beings must always be at the helm. AI is a tool to help with decision-making. Diverging points of view are nonetheless interesting: "AI is potentially more accurate, less biased and soon more creative than humans. Let's take advantage of it!" Along the same lines: "Humans can be corrupted. AI could have a stricter moral code than humans." Restrictions could also exist when an urgent decision had to be made.

Context is obviously everything: making croissants or launching an attack are obviously not the same thing. Humans should at the very least be able to make the decision to shut down an autonomous system. And that does not appear negotiable "in the case of complex decisions that include an ethical dimension involving responsibility".

SELECTED EXCERPTS

"There may come a point in system development where we will be able demonstrate that a human being can no longer make better decisions than a computer."

"Not the final decision, because the advantage of AI is to make a decision instantly based on a series of parameters that no human could ever analyze so quickly. But decision-making responsibility should always be assumed by a human being."

"Yes and yes, computer systems are there to assist us with decision-making, and that's how they must remain. Why give a cyborg control over us?"

"The fundamental decisions must be human and be as consensus-based as possible."

"You always want to have the option of an off switch. And we need to build systems in such a way that we can come to an understanding of how the machine is making the decision."

"Obviously with the current state of the technology, you can't let it have total control over too many things. That is unlikely to be true forever; eventually the AI is probably going to be smarter than the human ... and possibly more benevolent than the human, which is where you should really be putting your energy. At some point the question may be whether the human should even get any input into certain decisions, especially into decisions that affected more than just that human."

"If the human does NOT always make the final decision, then there needs to be a transparent interface so that users can correct the

decision-making computer system when it makes mistakes. (For example, with Google translate, you can provide a better translation.)"

JUSTICE (EQUITY, SOLIDARITY, DIVERSITY)

1. HOW CAN WE ENSURE THAT EVERYONE HAS ACCESS TO THE BENEFITS OF AI?

By making it affordable (or free), through open source and clearly exposing which decisions AI will be making for us (transparency). But is it possible in the capitalist system we know? "The private sector should not be able to exploit an annuity for its sole profit, at the expense of the rest of humanity." We could even tax companies that get excessively rich thanks to AI (would that harm innovation?).

Education could play a role in combating the digital divide. The role of governments (or even the UN) is to redistribute the benefits in equitable fashion and ensure that the AI values are aligned with human values. A basic income, a call for political realism tempers expectations: "let's not be utopian, AI isn't the one creating inequalities, humans are". One respondent also noted that information technologies make participative democracy possible. Another brought up the issue of basic income.

SELECTED EXCERPTS

"Build AI for the common good rather than private property. Regulate it to force advanced forms to adopt a GNU free licence for example, and promote information sharing."

"We cannot leave AI entirely at the mercy of the private sector."

"Possible AI advances, the discovery of a new protein for example, must be for the collective good."

"General quality of life for everyone should be improved with AI. Legal system seems to be one that will be greatly affected and see a lot of change, for the better."

"The digital divide could depend on whether or not large private companies get their hands on the data generated by the population."

"We must review international patent laws extensively. AI development will only truly progress if the information at its core is in the public domain. Special interest groups (corporations, army, governments) should not be able to appropriate this technology, otherwise it will inevitably be hijacked to serve their interests rather than those of the citizens."

"Make equity a central pillar. Include researchers and community groups that collaborate on designing equitable solutions. Take a look at the work done by the Support Unit (SRAP) and the mobilization and citizen involvement section of Alliance santé Québec."

"Give free Wi-Fi to the poor for starters."

"This is a very complex question. One could argue that everyone already benefits from AI through 'free' products like Facebook and Google Maps. What is missing is an understanding of the market value of someone's data relative to the machine's ability to build a more powerful model. Governments at all levels need to be using AI with the data they currently manage as another part of their policy-making tool set."

2. SHOULD WE FIGHT AGAINST THE CONCENTRATION OF WEALTH AND POWER ENJOYED BY ONLY A FEW AI COMPANIES?

Answers are clearly positive. By promoting open source and GNU free licences. Because it's the State rather than the private sector (GAFA) that citizens trust. The concerns are real: "Would democracy survive if predominant AI power fell into the wrong hands?" How could we even do this? We couldn't even manage to get open-source software to replace proprietary software. Nationalize to remain "masters of our own domain"? Regardless, AI should be seen as a common good that does not serve a minority. One respondent highlighted the need for an antitrust organization to break up certain monopolies. However, some preferred a more competitive model: "If some companies manage to carve a niche that brings them wealth and power, more power to them. But knowledge must remain in the public domain to cultivate competition." One respondent suggested that individuals own their own data and use an AI personal assistant that is loyal to them.

SELECTED EXCERPTS

"Obviously, we must fight the concentration of power, period."

"Yes, it seems there will be a lot of power available to those who control AI systems. New legislation/law will be required to monitor this, along with taxes on automation, etc."

"What's most important is that the basic programs are universal and built for the common good. Otherwise they will only be robots serving those who already maliciously rule the world with their own interests in mind, so either nothing changes, or the inequalities, violence, conflicts, etc. all get worse."

"The hands of a small number of AI companies or the hands of a small number human entities (i.e. the 1%) should not have more power and wealth than the 99% of human beings on earth. Powerful entities should adopt socially responsible behaviours at all time, especially when in presence of the public. (...) The democratization of AI should definitely empower the 99% of human beings."

3. WHAT KIND OF DISCRIMINATION COULD AI CREATE OR EXACERBATE?

Every "classic" form of discrimination seems to be exacerbated by AI, especially "social, racial, economic", but also "linguistic and cultural" discrimination, not only among people, but also among groups or states. A dystopian scenario looms: one where a new class of ultra-rich people (the 1%?) use AI to perpetuate socioeconomic inequalities. Participants also mentioned that access

to technology can be exclusive and excluding, especially for older people.

One respondent specified the type of mechanism that could encourage AI: "AI could be the perfect scapegoat under the guise of a BLACK BOX: Why didn't I get that line of credit, Mr. Bank Manager? Ah, I'm sorry, the system gave us that result." It also appears clear for respondents that humans, whether as individuals or as groups (e.g. systemic racism) that are responsible for this discrimination—not AI.

SELECTED EXCERPTS

"[We have to beware] of a 'caste' of AI experts emerging, whether known or secret, that holds all the knowledge, and therefore all the power. [We must also beware] of discrimination based on a health condition (flirting with eugenics), racial or sexual discrimination, towards the elderly, towards women, etc. [Lastly, also beware] of economic discrimination, increasing poverty for the majority and the power the rich hold over decision-makers."

"There are too many... That's precisely the problem. We have a hard time establishing what discrimination is or whether we're already doing it. How can AI determine this for us without discriminating exactly same way with the data we give it?"

"Social networks are already a source of stereotypes and racist, sexist, stigmatizing content. We can consider filtering that, offloading the problem. These filters could also be unduly discriminating."

"Algorithms must be developed by multidisciplinary and multicultural teams to avoid perpetuating stereotypes based on gender, wealth, race, etc."

"If AI contributes to well-being and autonomy, the people who need it, but can't access it, are worse off."

"If AI is deployed by special interest groups (armies, governments, corporations), it will only serve their momentary interests at the expense of the population."

"See weapons of math destruction. AI models with labelled training data that is discriminatory will simply perpetuate and reinforce these discriminations."

"It's going to be hard to deal with that, because in order to admit that the AI is going to find a regularity, you have to admit that the regularity exists. You have to be willing to say, 'Yes, XXX people *are* more likely to default on loans, but we want to ignore that anyway.' After that, it's a relatively simple technical problem to make AI implement your wishes. Short-term AI, anyhow."

4. SHOULD AI DEVELOPMENT BE NEUTRAL, OR SHOULD IT SEEK TO REDUCE SOCIAL AND ECONOMIC INEQUALITIES?

Most respondents are in favour of AI that would actively contribute to reducing social and economic inequalities. Many even consider this a priority. One optimistic respondent thinks that reducing inequalities will be the automatic result of AI development. Another wants it to mostly promote equal opportunities. However, a few skeptics would prefer it to remain neutral: "Who decides which inequalities to reduce?" And the more pessimistic maintain that there will always be inequalities … which shouldn't prevent trying to reduce them. Finally, one respondent suggests that AI remain neutral on social and economic inequality for commercial use, but that non-commercial use should aim for more equality.

SELECTED EXCERPTS

"Yes, that should always be its goal, along with reducing environmental impacts."

"AI cannot be neutral, so we might as well guide it in a direction that benefits everyone."

"Why are we developing AI? Reducing inequalities does not appear to be the primary reason; that does not mean, however, that AI development should be neutral: social and economic inequalities could serve as a 'constrained site', so the development doesn't occur at the expense of important values".

"AI models should be applied within a policy framework. No information system is neutral and any architect or policymakers must embrace the ethical challenges and opportunities when applying AI. In this context, reducing existing inequalities is a moral imperative. Machine learning models need to be conceived inside a larger pipeline that can mitigate regressions and provides recourse for error."

"But we should make sure that by doing so we are not actively causing friction between different groups or trying to homogenize them. The effect, in that manner, should be neutral."

"It should be neutral in commercial settings, otherwise the technology might never be adopted at all—leading to no benefit to the society. But it should also reduce socioeconomic inequalities in a non-commercial setting by giving everyone access to the same tools and opportunities."

5. **WHAT TYPE OF LEGAL DECISIONS COULD BE DELEGATED TO AI?**

The consensus was that no important decision should be delegated to AI. AI must simply serve as a tool to assist in decision-making. It could therefore "accelerate case processing", and even "make easy decisions after analyzing proof", such as decisions tied to paying tickets.

AI could be beneficial in other aspects of justice: "Detecting a lie or a false memory. Detecting risks of relapse." If a general artificial intelligence were developed, then AI that replaces judges could be envisioned; but this option is far from unanimous, even if it has been proven that human judges are

often biased in their rulings and subjected to various pressures. Perhaps the entire legal institution needs to be overhauled from top to bottom to make "artificial rulings" possible. Regardless, lowering costs and making justice more democratic would be good news and AI could certainly contribute to that, namely by making access to jurisprudence easier.

SELECTED EXCERPTS

"AI could replace a stenographer."

"AI would be fairer because it isn't subject to emotions or pressure from the media or other lobby groups. The only thing that would eventually need to be revised would be the Criminal Code, given the differences observed between human and artificial rulings."

"I don't believe any final decisions should be made by the AI. Seems the legal aid/technician and data processing could be best managed by AI."

"Decisions that involve complex practical rulings (jurisprudence) should be reserved for humans. Justice is also a social process. Let's not forget that.

"AI could do research for the general population (as well as jurists), by having access to all jurisprudence. This would make access to justice more democratic, since most of the costs for citizens go to jurists doing this kind of research."

"Current and near-future AI aren't going to be able to comprehend the law or apply it other than in cases so mechanical that you don't really need 'AI' at all. I suspect that any real legal decisions will take a truly general intelligence."

"AI predictive technology can be used to help judges make better decisions. The idea is not to replace judges."

PRIVACY (INTIMACY)

1. HOW CAN AI GUARANTEE RESPECT OF PRIVACY?

Many respondents wondered whether this question was relevant: How can AI ensure respect? The impression is rather that it violates it, repeatedly, without user consent. It also appears contradictory, since AI needs our data in order to develop.

But options may exist: "encrypt everything", not be invasive when requesting personal data. Someone remarked: "Respect of privacy is guaranteed if the person isn't exposed to AI by default." Users also have a responsibility: "It's up to each of us to control our exposure: shop in independent stores and pay in cash, rather than buying off the Internet."

Some were wary of the private sector: "Nothing is guaranteed if it's solely managed by the private sector." That is why we call the State and legislators to the rescue: Quebec's laws regarding privacy must be respected and improved. "It's a major challenge," because isn't it too late already? Our Facebook data, for example, may have been siphoned long ago by Cambridge Analytica or any other such company. And that's not even mentioning "hackers".

SELECTED EXCERPTS

"I believe the information economy, based on traceability, can create more information sharing, but at the same time more transparency in its use, and so having your information shared won't have such negative consequences if those who see it are also traced."

"Let's face facts, there are, realistically speaking, no truly reliable guarantees that AI can respect people's privacy. Health records & private accounts are hacked all the time despite the best security upgrades that technology has to offer. Google reads our private emails, doesn't it?"

"Differential privacy—the idea that you can give away information about yourself without ever having it trace back to you as the source. But, if such a practice is possible and can be made prevalent then I believe that informed consent is possible. The user has to be fully made aware of how the data being generated can and might be used, along with theoretical guarantees or open code base proving their claims."

"Make people's private data truly their private property."

2. DOES OUR PERSONAL DATA BELONG TO US, AND SHOULD WE HAVE THE RIGHT TO ERASE IT?

Agreement from participants was overwhelmingly positive for both questions. Someone specified, "and it should be very easy to do so, so everyone can do it". One respondent disagreed with the idea that our data belongs to us, but that that shouldn't prevent us from having "the right to examine its use". Although most respondents implicitly admitted that individuals should own their data, some see it rather as a collective good.

Erasing data should not obstruct justice (or healthcare services), which may need to access older data, nor should it harm others.

SELECTED EXCERPTS

"Yes, every citizen should own their personal data, just as artists own their creations."

"No, but data should be considered a national good, like libraries or nature reserves."

"Absolutely, and unequivocally. Only data required for good government functioning should be kept: demography, income, health, legal. All other data should be controlled by the user."

"As long as companies own and licence IP, individuals should have a right to all data they create."

"Generally yes. But I have a very broad definition of what should be considered personal data (and should be private property). Within this larger view even our criminal records would be personal data that we own (though not without

controls). It would be a category of personal data that we should not be able to delete—at least not whenever we choose."

3. SHOULD WE KNOW WHO OUR PERSONAL DATA IS SHARED WITH AND, GENERALLY, WHO IS USING IT?

The answer was a unanimous "yes"! Someone specified: "Just like we need to know who comes into our home, we need to know who can access our personal data." Another said: "Yes, [and we should know] who, how and for what." One respondent mentioned that we might grow tired of knowing who is using our data, and may quickly lose interest. But that obviously should not stop us from having the right to know.

SELECTED EXCERPTS

"Yes, I think I should even have a portal where I control 100% of the data I am sharing."

"Our data should never be shared without a clear and concise request to do so first being made. No 20-page contracts in small print where we have to guess that lifetime permission has been granted. If we subscribe to a service, information should never be able to be used in any other way than for the service requested."

"Absolutely and they should be required to ask permission to do so on a regular basis. Permission is not granted in perpetuity."

"Absolutely. Personal data should be private property. We should defend it and allow the owner to control who can access it and to what extent they can access it. The current default—wherein we cede our data to others—is bad for citizens and bad for democracy. There is another option."

4. DOES IT RUN COUNTER TO ETHICS AND ETIQUETTE TO HAVE AI ANSWER E-MAILS FOR YOU?

There were contradictory answers to this question. Many remarked that this kind of service already exists, or that certain people have human assistants that answer e-mails for them. One option would be to have AI prepare a response, but to have it validated by a human being (thereby giving them the last word). One respondent specified that "what's important, in my opinion, is that the person using this service has the necessary understanding and trust in the service". Another request that the process be transparent, meaning that the correspondent knows that the reply to their e-mail was written by AI. There may not be a generic answer to this question: it depends on the type of question ("Are you available for this meeting?" vs. "Do you think we should hire this person?").

SELECTED EXCERPTS

"No, as long as it is clearly stated that the response was written by AI rather than the individual concerned. If that person chooses to have AI answer for them, that's their responsibility … as long as the Internet service provider lets you activate or deactivate this feature. Obviously, it's not about imposing this service.

"It's useful for those who have to manage a high volume of similar messages with low complexity."

"That depends, if you always answer the same way to the same questions, it's not going to make much of a difference for you."

"If it's a question of customer service, of fulfilling a human need to satisfy that involves responsibility for others, I expect a human being to answer."

"Yes. Human intent is a critical component to our society's framework. We can delegate to AI, but human dignity claims that you should know if you are interacting with a machine."

"Similarly, if an organization has a bot deal with people, it should always identify itself as a bot. People should always know if they are dealing with a human or a machine. And the organization that has bots dealing with people should always be held responsible for any actions the bot takes on the organization's behalf."

5. WHAT COULD AI DO ON YOUR BEHALF?

An open question that elicited very different answers, ranging from "nothing" to "everything" (as long as consent was provided). Between the two extremes: book an appointment manage my finances, my schedule, file my taxes and other administrative tasks, vote (!). But I should always be held responsible for the consequences of what AI does on my behalf. (Many respondents confused this question with "What could AI do for you?", e.g. vacuum.)

SELECTED EXCERPTS

"Everything that I have approved beforehand."

"Any task that does not commit to any future engagement."

"Nothing serious that could have legal or emotional repercussions."

"Book appointments respond with numerical data that is already in the public domain, check on the well-being of family pets."

"That depends on the AI. I wouldn't trust any *present* AI to do anything that I couldn't countermand or that people would interpret as a direct application of my personal judgment."

"My recommendation is to adopt a paradigm in which each citizen owns private, 'loyal' AI tools (agent) that can help protect, manage, analyze and use a citizen's private data (stored in a protected online profile) to help that citizen at their behest and only their behest. (...) Some people might say they can do simple repetitive tasks, perhaps review email. Others might allow their AI agent to browse the web to plan online shopping. Others might let the agent actually make purchases autonomously. Others might allow the AI agent to perform investment transactions for them. In an advanced future, some prefer

to trust their AI to participate in a family vote about 'pulling the plug', given on its intimate access to its owner's private data, which could analyze a variety information types taken from a personal profile, allowing it to use predictive analysis to help decide what the citizen might want if they were able to speak."

KNOWLEDGE (PUBLICITY, CAUTION)

1. COULD AI DEVELOPMENT JEOPARDIZE CRITICAL THINKING?

Answers varied, but leaned towards "no". On the "risk" side, many fears are raised: loss of sense of curiosity, publicity, standardizing thought and dismissing marginal viewpoints. AI might also speak on behalf of humans and appear too reliable: "The machine can't be wrong; everything has been said, there's nothing left to add".

On the plus side, many noted that the time gained through automation could be invested in critical thinking and the fact that AI and information technologies make information more accessible, or that we could even program AI to perform critical thinking— the idea that AI could be more neutral than humans was also brought up. Finally, AI could be viewed as a wonderful opportunity—or need—for humans to exercise critical thinking.

SELECTED EXCERPTS

"Yes, but not if it is used to make people's lives easier thus leaving them with more time to educate themselves and develop their critical thinking."

"No, quite the opposite. The sum total of human knowledge is growing at an exponential rate, to the point where it has become impossible to know all the ins and outs of a problem. AI, with its ability to summarize, allows humans to filter redundant information and focus on what's essential."

"I believe it certainly could compromise humans quest for knowledge & need to problem solve & therefore seriously impair our critical thinking & problem solving capacities & increase depression in people who may in future, have no motivation to use their god-given gifts & intelligence because they have been replaced by AI."

"It would definitely be more of a crutch than a tool if we become overly reliant on it. Instead the development and the products that are created using AI tech should be such that it aids critical thinking, aids skill development and indirectly making life easier."

2. HOW CAN WE STOP FAKE NEWS OR FALSE INFORMATION FROM SPREADING?

This question was open-ended and generated a wide range of potential solutions: providing financial support to media (local, traditional) that fact-check information, investing in quality journalism (with multiple information sources), educating people, using AI to fact-check information, punish those who spread false information, erase it, impose regulations for platforms (such as Facebook) that spread these

fake news. Our collective dependency on "free" (one-way only) news was also highlighted.

Should fake news be censored? One respondent took a stand: "We should circulate fact-checking articles as much as possible instead, because censorship is counterproductive (it can feed conspiracy theories, for example)". One pessimistic point of view: "It may become impossible as AI advances so too will its ability to mimic voices and fabricate images and video."

SELECTED EXCERPTS

"Redefine the journalism profession. Develop an accreditation system for information sources. Recognize communication experts in various sectors of human activity."

"There will always be fake news, we must develop critical thinking and educate youth on the matter."

"Censorship must not come directly from AI. However AI can become a tool to help predict the likelihood of a news item being fake."

"Teach people how to develop critical thinking, search for credible information and open their minds."

3. SHOULD THE RESULTS (POSITIVE OR NEGATIVE) OF STUDIES ON AI BE MADE AVAILABLE AND ACCESSIBLE?

The answer was positive beyond a shadow of a doubt. Many respondents felt that this should be the case for study results in all fields. These results should be open source, according to other respondents (it should be noted that a vast majority of them already are).

SELECTED EXCERPTS

"Absolutely. And as much as possible, break down these results to make them accessible to all. No opaque results, with incomprehensible terms..."

"This question has more to do with research than AI. Publicly funded research, with few exceptions, should be made available as a Social Good."

"Yes. I know people who think really powerful results should be kept from the 'bad guys'. That is a total pipe dream. All you'll do by trying is to disadvantage the 'good guys'. Your best bet is to be open."

"YES!! Especially negative results. They would provide as much information, if not more about a particular problem."

4. IS IT OKAY NOT TO BE INFORMED WHEN MEDICAL OR LEGAL ADVICE IS DISPENSED BY A CHATBOT?

For our survey respondents, the answer was predominantly "no". Their answers were influenced by two concerns: transparency and privacy: "Advice dispensed by a chatbot may be taken into consideration differently if the person knows they're speaking with a chatbot, or believe they're speaking to a human. A chatbot cannot know all the variables for a given situation." Many mentioned that it was easy to let a person know that they are speaking with a chatbot.

SELECTED EXCERPTS

"Eventually, yes. No passenger on a plane asks the M/C whether it's the pilot or the autopilot who is steering the plane."

"The source of such advice being often critical to a person's well-being, one should be aware of the source of this information."

"No, all information should be presented along with the source exactly as it is, along with the analysis of how accurate or biased the information/advice might be. It may happen that the person may rely on that information even after realizing that it is from a chatbot, as it would get good results. And that is the kind of relationship we'd like to foster."

5. TO WHAT EXTENT SHOULD ALGORITHMS BE TRANSPARENT ABOUT THEIR DECISION-MAKING PROCESS?

This question left many respondents uncertain. The most popular response was "as much as possible", while acknowledging the technical difficulties in play (the "black box" problem). Although some believe that AI should simply not make any decisions, others seemed to agree that AI can make decisions, on the condition that there is access to a "justification that can be understood by a human". Transparency may not be desirable in certain contexts. Many mentioned that transparency is important in building trust in AI. One respondent suggested giving a reliability rating for decisions made by AI.

They also noted that transparency involves knowing which data (or which type of data) an AI makes its decision and the values (or interests) guiding its decision.

One participant suggested instead that we not ask any more from AI than we would from a human.

SELECTED EXCERPTS

"A description of the algorithm's decision-making process should be included with the purchase of an AI product, like an instruction manual or the manufacturer's warranty that comes with regular purchases."

"If AI creators cannot precisely define the reach and the limits of an AI's decision-making ability, then that AI shouldn't be marketed and sold."

"The scale of values used to make their decision. See the relative values for different decision-making elements. For example: cat vs. dog, collective vs. individual, etc."

"Completely transparent. How can you trust something if you don't know what principles they are basing themselves on to conduct their analysis? Just like understanding the methodology used by researchers is always relevant."

"You should be able to ask AI why it made a choice then if you find its reasons lacking you should be able to make it change its behaviour."

"We may be able to infer decision-making processes but we should not assume that there is any internal motive or intent in an algorithm."

DEMOCRACY (PUBLICITY, DIVERSITY)

1. SHOULD INSTITUTIONS CONTROL AI RESEARCH AND APPLICATIONS?

The response was positive overall, especially for AI applications (freedom of scientific research is an important value). An "office of the AI Ombudsman" was suggested, along with AI ethics committees or some sort of Hippocratic oath. Participants also noted that "the subject is too intensely political and social to be left in the hands of the private sector". This control, however, should not impede innovation (as long as it is compatible with the common good and human rights). One inherent difficulty for institutional control stems from international politics: how can countries with competing interests agree on common institutions?

SELECTED EXCERPTS

"Yes, on the condition that we develop a participative democracy and that governments are in the service of the majority, not of money."

"No, but establishing boundaries is essential."

"Yes but good luck getting China or Russia to follow along."

"Controlling AI research is simply not possible. The research itself should continue, but a broader communication framework explaining what AI can and cannot do is critical. Sensitizing researchers to the ethical ramifications of their work is also important (e.g. the Hippocratic oath).

2. IN WHICH FIELD IS THIS THE MOST PERTINENT?

The question was open-ended. Many answered, "in all fields". Healthcare easily takes the lead in the fields listed, followed by weapons, justice, environment, food, surveillance, privacy, finance, safety, education and government, respectively. The following were also mentioned: economy, industry, epigenetics, journalism, transportation, municipal services, research on a super-IA (AGI), self-driving cars and targeted advertisements.

SELECTED EXCERPTS

"In all fields related to life (biology) and living in society."

3. WHO SHOULD DECIDE—AND WHAT SHOULD THE TERMS BE—ON THE STANDARDS AND VALUES DETERMINING THIS CONTROL?

Respondents were often unsure how to answer to this, and hesitated between various options: Parliament, public consultations, the overall population (referendum, random draw), a multidisciplinary committee (experts, elected officials, citizens), a science and technology ethics commission, an advisory committee, an international institution (UN-style). The idea that this decision-making body must be independent (from political and economic power) was raised many times, along with the concern that this body must represent the diversity of citizens.

SELECTED EXCERPTS

"I don't know... A joint, multidisciplinary, academic, public and impartial committee."

"All of us, by developing information resources, consultation and decision-making methods that involve as many people as possible from all walks of life. Not the current "democracy".

"A lot of committees. They could establish rules, values, etc., tied to each institution where there would be one of these committees. They could thereby establish some sort of "charter" for the institution and make recommendations... That obviously shouldn't be left to gather dust on a shelf!

"In Quebec, the science and technology ethics commission already produced a document on smart cities outlining the issues to consider. Other AI projects could be analyzed by this body or other government bodies specializing in the field. An ombudsman could be named to certify AI projects and receive flags about the Montreal AI Declaration principles not being respected."

"Since AI affects every field (law, health care, science, society, arts), specialists from each of these fields must be represented within the organization. The government must fund this organization properly, but cannot intervene in how it operates. Furthermore, the government should not have the power to eliminate the organization or interfere with its work."

"Canadians from all groups, backgrounds & beliefs."

"This should function like an IRB as in the drug development and testing industry."

4. WHO SHOULD CHOOSE THE "MORAL SETTINGS" FOR SELF-DRIVING CARS?

There were a number of different answers to this question: Parliament, a government agency, the State, provincial powers, the State in collaboration with the industry, an ethics expert committee, the SAAQ, the car manufacturer, a software certification authority, a user committee, Supreme Court judges, a U.N.-like international organization. The user could also have the choice of certain options. It should be noted that many respondents distrusted self-driving cars ("they should be banned").

SELECTED EXCERPTS

"Again, it could be citizen committees. We'd need a representative for pedestrians, one for seniors, another for youth 16 and under, another for bikes, etc. Each one could have a say on the moral settings for self-driving cars."

'Certainly not the companies building them!"

"An ethics commissioner and the Bureau du Coroner in Quebec."

"It should be a multilateral decision (after thorough public discussion)."

"Judges/supreme court, whoever decides and upholds the existing ethical guidelines should have a major role to play in the decision. But along with them, community participation, transport businesses and authorities, AI researchers and developers."

5. SHOULD WE DEVELOP ONE OR MORE "ETHICAL LABELS" FOR AI, WEBSITES AND COMPANIES THAT RESPECT CERTAIN STANDARDS?

The vast majority of participants agreed, saying it was a "good idea", a "good start". It could be similar to an ISO standard. One respondent wondered, however, why all companies and websites did not have to respect these standards. Another specified: "yes, case-by-case with a standard chart". This also raised some skepticism: Will these certifications be respected? Don't they risk being corrupted?

SELECTED EXCERPTS

"Certifications that would eventually be subject to a vigilant review to adapt to a given situation."

"Communities are different, people are different. (...) We should make sure that by doing so we are not actively causing friction between different groups or trying to homogenize them. The effect should be neutral."

"Definitely, at least three major ones should be developed: corporate, government and individual ethical labels."

RESPONSIBILITY (CAUTION)

1. WHO ARE THE STAKEHOLDERS RESPONSIBLE FOR THE CONSEQUENCES OF AI DEVELOPMENT?

Respondents identified numerous stakeholders: universities, researchers, companies, ethicists, politicians, those who market the apps, the government, the economic decision-makers, those who hold a financial stake, elected representatives, society, users, every one of us. But it was probably "the developers/creators, the companies and the government" that came up most often. Some drew a parallel with pets or children: the owners/guardians are responsible. In the case of AI, it could be the owners, or even those who test the AI, who authorize its deployment.

SELECTED EXCERPTS

"The people who build them, the people who distribute them, and, if we can nab them, the people who use them maliciously to harm, injure, kill or dominate others (including animals), or harm the environment."

"Every member of the supply chain: from the graduate researcher to the multinational firm, including local, regional and national regulatory bodies."

"Companies offering services must be accountable and responsible, but especially company stakeholders."

"Whoever provides the results/ predictions of the AI decision-making. For example, Google is responsible for Google Translate."

"Researchers developing models are partially responsible. However the application of AI ultimately rests with the owner/operators."

2. HOW CAN WE DEFINE PROGRESSIVE OR CONSERVATIVE AI DEVELOPMENT?

Participants had no definitive answer to this question. Progressive development is synonymous with the collective, transparency, smaller wage gap. Conservative development goes hand in hand with caution: there's no point rushing in, better to go gradually. Someone remarked that it seemed easier to adapt legislation around AI than adapt AI around legislation because progress moves quickly and shows no signs of stopping. Another mentioned that progressive development should "Foster alternative research". And a sentiment shared by many was: Let's go, let's go, can we do things differently?"

SELECTED EXCERPTS

"By holdings forums on the subject! :-) The more we talk about it, and the more inclusively, the more progressive AI development will be, in a good way. Also through education. The more our society is educated, the more informed it will be, the more informed its decisions will be."

"For common good vs. private property."

"It is progressive when it is maximizing freedom and agency. It is conservative when it is carefully monitored and cultivated as to insure safety."

"Conservative development: Checking, testing at each and every step. First in isolation, then within an isolated test group, and gradually deploy the AI."

3. HOW CAN WE RESPOND TO PREDICTABLE OUTCOMES FOR THE WORKPLACE?

Many ideas came back over and over: a solid social net or basic income, a tax reform with a tax on robots, or a better distribution of wealth. Access to education and training is the preferred route; however, people will have to adapt, which requires more ongoing training. The transition will certainly have to be gradual and transparent: people must be kept informed. But not everyone is worried: The workplace has always been evolving and will continue to do so." Incidentally, many seem to hope to free themselves from work.

SELECTED EXCERPTS

"Offer a guaranteed salary in exchange for participating in the creation of digital commons."

"Work is not humanity's ideal, nor is it its goal. The free time obtained and the productivity gains generated should be pooled to allow everyone to work less without sacrificing standards of living."

"By redirecting people towards other types of employment that are more involved in building social cohesion."

"Need to slow down the pace; we must first define priorities that aim to develop what can serve humans before what can replace humans."

"AI tax, job displacement compensation, basic living wage, and research/development of new jobs."

"The real cost of the introduction of AI technology is not just the money some people pay for it. It is the social, political, and economic costs—to everybody in society that need to be considered."

4. IS IT ACCEPTABLE TO ENTRUST AI WITH CARING FOR A VULNERABLE PERSON? (FOR EXAMPLE, WITH A "ROBOT NANNY")

Respondents were very torn on this question: "to entertain, but not to heal", "not sure". There seemed to be a fear of humans in healthcare disappearing. The importance of "human warmth", especially for vulnerable people, was brought up. Of course, it still seemed better than nothing: Yes, if there's no other choice." There is also the fact that it may provide better access to care, especially when human resources are scarce. Many highlight the risk of shirking our duties toward these people by entrusting them to AI. The subject is a sensitive one and such robots should certainly be guided and supervised.

SELECTED EXCERPTS

"Not completely. The robot nanny should always be there as a complement to human staff."

"Yes, if you can program AI correctly so it doesn't bypass certain more sensitive skills."

"Up to the vulnerable person to decide."

"No. The result could be disastrous as it has not been studied for decades to determine the social, psychological, mental & physical effects it would have on our children. It could also possibly make our children emotionally unable to connect & bond with their parents, siblings & other humans."

"Of course ... consider how television is sometimes referred to as a babysitter."

5. CAN AN ARTIFICIAL AGENT SUCH AS TAY, MICROSOFT'S "RACIST" CHATBOT BE MORALLY BLAMABLE AND RESPONSIBLE?

The question drew mostly negative responses. The chatbot is not defined as racist "because it doesn't understand anything", and the blame is rather placed on its designers (Microsoft). Nonetheless, the "consequences of its declarations" could have very real impacts. Most respondents therefore agreed that it is unacceptable. One respondent brought up the legal aspect by considering placing AI under guardianship (like children or animals), while another considered them simply as objects for whom responsibility falls on its owner.

SELECTED EXCERPTS

"No, I think we should consider artificial intelligence products as if they were children. Giving them the title of a person without a complete legal personality would be a good idea. That way, each artificially intelligent product would have a human guardian that would be responsible for its actions."

"In the end, it's just a program. And we know to what extent some programs can be bugged, faulty and poorly made."

"Not for the moment, responsibility comes with sentience, if AI isn't sentient, it can't be blamed."

"It's the programmer's responsibility to make sure its robot isn't racist and to make any required changes as quickly as possible."

"We should accept that machine learning algorithms are non-deterministic and empower operators to explore their utility while being responsible operators."

"The responsibility (until proven that the being is actually sentient, if that's even possible) should be taken by: People who gave permission to deploy them > People who tested them > People who developed them. In that order."

"Humans are not good examples for AI agents. AI agents will be more efficiently learning from other AI agents than from human activities."

"No. I think it is always people who must be held responsible. I am against giving machines any kind of legal status similar to people. You cannot punish or hold responsible a machine. So, people must always be responsible."

3. SUMMARY OF SUBMISSIONS RECEIVED

Over 15 documents were submitted following the call for proposals published on the Montréal Declaration's website in November 2017 (with a deadline at the end of April 2018). The objective of this was to contribute to the Declaration's content, either by discussing the seven principles in the preliminary version, or by suggesting concrete recommendations. These documents range from summary reports of collective discussions to individual opinion pieces. They are written in French and English, and can be read on the Declaration's website (this summary obviously cannot do justice to the rich content of the submissions received).

The following abbreviations are used to indicate the documents from the following people or organizations:

AQT
for Association québécoise des technologies

CAIQ
for Commission d'accès à l'information du Québec

MAIEM
for the Montreal AI Ethics Meetup group

OIQ
for Ordre des ingénieurs du Québec

SRAD
for the evening of reflection around the Declaration which was held at UQAM

Hernandez
for Annick, Guillaume and Raphaël Hernandez

McNally
for John McNally

Musseau
for Pierre Musseau-Milesi

Parent
for Lise Parent

Quintal *et al.*
for Ariane Quintal, Matthew Sample and Eric Racine

Ravet
for Jean-Claude Ravet

Robert
for Bruno Robert

Wark
for Grant Wark

PRIVACY

PROPOSED PRINCIPLE

"AI development should guarantee the respect of privacy and allow people who use it to access their personal data as well as the kind of information involved in the algorithm."

GENERAL OBSERVATIONS

The privacy principle has probably been commented on the most in the submissions received. The Commission de l'accès à l'information du Québec (CAIQ) in particular, but also the Montreal AI Ethics Meetup (MAIEM) group, the discussion session on the Declaration held at UQAM (SRAD), the Ordre des ingénieurs du Québec (OIQ), Lise Parent (Parent), Annick, Guillaume and Raphaël Hernandez (Hernandez), Grant Wark (Wark), Quintal, Sample and Racine (Quintal et al.) all suggested recommendations explicitly linked to privacy.

As the CAIQ mentions, in Quebec, there are already well-established principles for the protection of personal information (RLRQ, A-2.1 ; la Loi sur l'accès, as well as RLRQ, P-39.1 ; la Loi sur le privé) that AI development will have to respect: for example, the organizations collecting data must determine ahead of time the reason they are collecting this data and advise the people concerned. Once more the principles of necessity, consent, confidentiality, destruction, transparency, access and responsibility (see CAIQ appendix) can be noted.

Regarding new practices, at least two types of regulation can be considered: one coercive, which focuses on penalties in the event the legal framework is not respected, and the other preventive, which aims for greater flexibility in adapting to change. In Quebec, the CAIQ suggests the second approach and insists on evaluating the risks beforehand, using parameters with the strictest possible default settings, using technology to improve confidentiality, designating a person in each organization who is responsible for the protection of personal information and held accountable as well as "transparency, working for citizens". We have to wonder, however, if the balance of power with major digital multinationals will not also entail more coercive than preventive measures.

This position perhaps echoes that of the OIQ (and Parent) which promotes privacy through design and suggests drawing inspiration from existing best practices, such as the General Data Protection Regulation (GDPR) which recently came into effect across Europe.

This concern for the respect of privacy is often accompanied by a concern for transparency. The MAIEM group suggests, therefore, expanding on the privacy principle by specifying that transparency is essential—an analysis also made by the CAIQ and the SRAD. The close relationship between the issues of protecting certain information (personal data) and being able to know who holds what (access to information) is evident, two elements which are likely to be expanded in the Declaration. We also note the tension that sometimes surfaces between these two elements: when transparency applies to personal information that we would rather keep confidential. Mediation between these two notions may prove necessary.

As well, consenus on this mediation may not be reached, because as the MAIEM highlights, privacy preferences can "vary considerably according to cultures, generations and individuals". One idea for which there is certainly consensus it that we must "preserve citizen control over their personal information and the management of their consent" (CAIQ, SRAD). Quintal et al. also worry that the initial formulation of the privacy principle suggests that data be shared by default (the principle insists on being able know what becomes of personal data, without objecting to the data collection in the first place). "The Declaration should include improved safeguards for privacy of user data."

Lastly the SRAD notes that data anonymity techniques are not yet mature enough to be used without risk. the SRAD also observes the link between data protection issues and the risks

of algorithmic discrimination. But that does not mean that protected data (for example gender or race) should not be collected insofar as fighting discrimination usually implies having access to this kind of information.

SUGGESTED RECOMMENDATIONS

The privacy principle, which includes the concern for transparency, leads to more specific recommendations:

> People need to be informed of, authorized to and able to check use of their personal data at any time (MAIEM).

> We must introduce a culture of "data privacy by default" as is the case with neuroethics, meaning that by default, personal data should not be shared (Quintal et al.).

> The "burden of consent", meaning ensuring that consent is truly free and informed, should fall on companies/organizations and not citizens, just like correcting erroneous information (CAIQ).

> People need to be able to understand how their personal information is being used (MAIEM, CAIQ, Hernandez, Parent).

> People must be able to withdraw their consent regarding use of their personal information (MAIEM).

> Computer codes for interpreting results and algorithm training methods must be made public and open. (OIQ)

> We must make people aware of privacy protection issues (CAIQ).

> People should be able to know the monetary value of their personal information at all times (Hernandez).

Lastly, an original and detailed proposal from Wark answers, to a certain extent, a question put forward by Hernandez : How can we create a private digital space? Basically, it is a matter of using AI to protect against AI.

> Indeed, Wark suggests using "smart contracts" technology to protect personal information and make business transactions and social interactions easier. This can be achieved by developing a secure personal profile and a "loyal AI" that would serve as a personal data manager, thereby solving many of the challenges previously identified. "For example, a loyal AI-agent must not compromise its loyalty to its owner through overt or covert association with a business, such as an online store." To find out more, refer to Wark's document which gives a detailed presentation of what loyal AI might look like.

Many papers discuss how these recommendations can be implemented. From a public policy standpoint, there are at least three ways to respond to this concern for respect of privacy and transparency: through regulation, self-regulation or incentives.

Both the CAIQ and the MAIEM agree that self-regulation is insufficient. Updating existing regulations is more important. Both organizations (as well as Parent) also stress the importance of conducting business and organizational audits. This update could go in a different direction: OIQ supports "flexible regulation mechanisms", which aligns with the preventive approach adopted by the CAIQ.

Finally, we can consider financial incentives for companies that develop technologies to protect privacy, and promote those who make efforts, namely through labels or certifications—a sentiment that seems to be shared by the Association québécoise des technologies (AQT).

JUSTICE

PROPOSED PRINCIPLE

"AI development should promote justice and seek to eliminate discrimination, especially when it comes to gender, age, mental and physical capacity, sexual orientation, ethnic and social origins and religious beliefs."

GENERAL OBSERVATIONS

Like the privacy principle, justice was also present in many of the proposals: MAIEM, SRAD, OIQ, Hernandez, Parent, McNally, Ravet.

The SRAD suggests distinguishing between the various meanings of justice (according to Aristotle): commutative justice that oversees exchanges between people who are considered equal, and distributive justice which is linked to merit. Who deserves what in society? This second meaning is the one that appears mobilized in the submissions received, and it raises a number of questions.

Is it possible to identify a universal justice principle to regulate AI development? Should it not be limited to principles that apply only to a given community? This delicate issue lies at the core of many political philosophy debates.

The MAIEM leans towards a non-universal approach, or at least one which tries to take cultural and historical variations on the notion of justice into consideration:

"The development and utilization of AI-enabled solutions should promote justice and human agency as transparently defined by the target community's welfare-defining organization (e.g. democratically elected government), in concert with the target community. It should seek to eliminate inequality and discrimination within that community."

The counterpart to this reformulation exists in Ravet's more universal approach, which identifies a universal principle in Kant's idea of human dignity and life: "AI innovations must be based on the principle of non-instrumentalization of humans and be careful not to crush life." This approach is also favoured by the SRAD who, in addition to the notion of equal dignity of human beings, introduces the idea of social justice: "AI development should promote social justice and respect equal human dignity, particularly by seeking to eliminate all forms of discrimination especially with regard to gender, age, ethnic origins, social status, etc."

One way to articulate social justice and justice as non-discriminating would be to see the first as correcting (socioeconomic) inequalities, whereas the second seeks to prevent inequalities from appearing and guarantees equal chances. Social justice can also be considered in greater context, as the MAIEM does when it underscores the need to consider different perspectives on justice, especially those from marginalized communities.

SUGGESTED RECOMMENDATIONS

The question of biases (already discussed in the previous section) and the opacity of algorithms (the "black box" problem) also caught people's attention. This is unsurprising given that the issue has received a great deal of media coverage. For example, Parent notes that "assisted, or even automated, decision-making systems in medicine, finance, defence or justice, will give biased results if their input is biased". The OIQ also insists on the need to implement "control and protection mechanisms" to correct the bias.

Other recommendations are also worth mentioning:

> We must train students and AI practitioners in law and ethics. (Parent, OIQ)

> We must foster diversified and female hires in AI system development. (OIQ)

> We must ensure quick and transparent processing of claims by users/citizens who have been negatively impacted by an AI system (OIQ).

Many proposals call for the creation of an **independent oversight body** (Parent, McNally, Hernandez, OIQ, AQT). Its role would not be limited to applying the justice principle, but as it often appears in discrimination issues, this a good opportunity to discuss it.

The form this will take varies from one document to the next. The OIQ talks about an AI observatory, Hernandez describes, "a regulatory body whose task would be to ensure that citizens have a good understanding of the decisions made by AI"; as for the AQT, it advocates "the implementation of a multisectoral advisory committee whose purpose would be to reflect on the opportunities and challenges for Quebec's technology industry in the matter of ethics in artificial intelligence". One could also envision, as McNally does, an oversight organization that would work closely with the government and whose mandate would be to anticipate problems that AI might cause for the society of tomorrow.

RESPONSIBILITY

SUGGESTED PRINCIPLE

"The different stakeholders in AI development should assume their responsibilities by working to counter the risks of these technological innovations."

GENERAL OBSERVATIONS

The responsibility principle does not appear as often as the previous two in the submissions received, but it tends to overshadow the question of the relationship between humans and AI. Who will be responsible for AI, especially its adverse effects? As the SRAD remarked, AI development could extend all the way to using killer robots. This possibility raises, in turn, a commonly shared concern: that humans are handing their responsibilities over to AI. Here we find the theme of AI as a tool: it should be viewed as an extension of human intentionality, but not as an autonomous intentionality (MAIEM).

Among the people and groups responsible, we can include the researchers who, because they possess the knowledge, must start the debate (SRAD). To this we can add those who sponsor the researchers, such as universities, the military or the industries. Being responsible namely means implementing the knowledge and tools to "understand the functioning of AI and anticipate its reactions" (MAIEM).

In an essay that offers a broader outlook on the prevailing understanding of AI rather than expanding any specific principles of the Declaration, Jean-Claude Ravet, editor in chief of the magazine *Relations*, cautions against human instrumentalization in the age of AI and believes that AI development is our collective responsibility, and that we must maintain a global perspective that is both historical and ideological. Thereby, the motive itself of AI as a tool is worth questioning, since "the line between using the technique and the technique itself is blurrier than ever". Most importantly, notes Ravet, we should not kid ourselves about the ideology behind AI development that serves the interests of powerful multinational corporations. For Ravet, this ideology, which tries to pass itself off as scientific speech rather than a social project, is characterized by "an extremely reductive vision of humans and life". (Hernandez also questions the specificity of humans).

The transhumanism movement or the book *Homo Deus* by Yuval Noah Harari are good examples of this reductionist ideology that Ravet condemns: "Under the pretext of making humans more, we must not make them less and make it a means to an end. The sole criteria of making money isn't enough. Nor is the respect of individual choice. Because the issues affect life and humanity itself." We need to look critically look at what often appears obvious: that humanity progresses because of AI and that it is inevitable these machines will make their way into our lives. In other words, we are collectively responsible and that is why humans must always have the last word "as beings capable of speech, feelings, sensations, who are aware of human fragility and the ties that bind them to others, to life and to the Earth" (Ravet).

SUGGESTED RECOMMENDATIONS

- Human beings must ultimately be held accountable for AI-assisted legal decisions (SRAD, Parent, Ravet).
- In the case of engineers, we must ensure professional accountability (OIQ)
- From the standpoint of legal responsibility, we must anticipate eventual disputes over AI systems with non-Canadian jurisdictions (e.g. components designed or built elsewhere than where the system was used) (OIQ).
- To avoid attributing undue responsibility to AI, they should not have the misleading appearance of a moral patient (meaning an individual that can be wronged) that deserves our empathy (MAIEM).
- The formulation of the principle, the intention to "counter the risks" does not go far enough: the people responsible must assume the results of AI development (MAIEM).

WELL-BEING

SUGGESTED PRINCIPLE

"AI development should ultimately aim for the well-being of all sentient beings."

GENERAL OBSERVATIONS

Like responsibility, the well-being principle is often present implicitly, especially in health, safety or even the equal distribution of AI benefits. In fact, according to certain approaches in moral philosophy, this principle could even serve as general criteria for decision-making: when given the choice, we should act so as to create as much well-being as possible. Obviously, as the MAIEM observes, other values may conflict with well-being, especially autonomy. For example, we can find situations where paternalism seemed justified, such as when a moral patient's autonomy is constrained for their well-being. It is hardly surprising, then, that such conflicting values—often discussed by philosophers in moral dilemmas—are considered in the submissions received. However, it is nonetheless true that a principle on well-being must be simple, easy to understand and leave some room for future interpretation (MAIEM).

For its part, the OIQ states that the well-being principle is aligned with one of the main tenets of the engineer's code of ethics (article 2.02) which stipulates that the "engineer must respect their obligations towards mankind and take into consideration the consequences their work will have on the environment and on everyone's life, health and property". For this reason, promoting well-being implies evaluating, to the greatest extent possible, risks related to the deployment and operations of AI applications, keeping in mind that "there is no such thing as zero risk" (OIQ).

We should note that the very inclusive character of this principle, which not only targets the well-being of humans, but of sentient beings as a whole, was not questioned. It may be a sign of our changing mentalities and our relationships with non-human (sentient) animals. The MAIEM and the SRAD expand their notion of the domain of morality to sentient beings while Parent brings up AI interference with animal life. We also note that some papers (Ravet, MAIEM, Parent) seem interested in considering the criteria of life and extending the circle of morality to non-sentient entities (such as plants and ecosystems). These concerns, however, which could be qualified as biocentric, have not been adequately developed to be said to reflect a (fairly radical) moral position: it may be a concern for an anthropocentric environment.

Ideas also seem divided on whether the capacity for AI to be sentient (or sensitive) would be adequate criteria on which to grant it rights or, at the very least, moral consideration. If, for example, a robot could suffer, it would have a legitimate interest in being protected. This point remains highly speculative since AI systems are still very far from having feelings or emotions.

Lastly, in a somewhat speculative and programmatic text, Museau attempts to articulate the notion of moral minimalism developed by the philosopher Ruwen Ogien and the well-being principle recommended by the Declaration. What stands out is that the goal of AI development should be to not harm others nor to improve itself—self-improvement belonging, according to Museau, to both moral maximalism and transhumanism.

SUGGESTED RECOMMENDATIONS

> Formulate the principle to reduce suffering rather than promote well-being (which corresponds with what is sometimes called negative utilitarianism (MAIEM).

> Out of safety concerns, blocking/disengagement devices must be planned when designing AI systems to maintain control in case of failure (OIQ).

AUTONOMY

SUGGESTED PRINCIPLE

"AI development should promote the autonomy of humans and responsibly control that of computer systems."

GENERAL OBSERVATIONS

With regard to autonomy, consensus was reached on promoting human autonomy. This idea especially translates into the theme of AI at the service of humans, as mentioned earlier. The OIQ notes that "robots and AI systems must be seen as tools to assist or help with decision-making, not as a replacement for human judgment". For his part, Ravet insists that humans should not be reduced to machines nor become a means to an end, while Hernandez wonders if AI won't one day replace humans to the point that they become obsolete.

Nonetheless, the truth is that the idea of autonomy is subject to multiple interpretations. The SRAD proposes a detailed analysis grid of the types of autonomy ("condition of an entity which chooses itself the rules to which it submits") divided into moral, political and functional (non-dependence) autonomy. These three types of autonomy can be cross-referenced with three types of situations: the autonomy of a human assisted by AI (for example a person with a disability), the autonomy of a human in an environment populated by AI, and finally the autonomy of AI in a human environment. The SRAD suggests a reformulation, therefore, that further considers these diverse meanings: "AI systems must not harm the autonomy (moral, political and functional) of human beings, but rather seek to contribute to it. AI systems must not be made entirely independent of human beings, but remain under their control (moral, political and functional)". This being said, we should not jump to the conclusion that autonomy should systematically prevail over other values such as well-being, justice or knowledge. Each case must be examined in context. And as the MAIEM reminds us, people's consent remains a good way to guarantee their autonomy.

Although there was consensus on the value of human autonomy, the issue of "AI system autonomy" was more sensitive in that its guardianship could be contested. Therefore, citing an article on digital evolution and artificial life, the MAIEM reminds us that situations where the autonomy and creativity of AI systems could contribute to the general well-being are foreseeable. Nonetheless, the MAIEM states that the autonomy of an AI system should not be sought out in itself if it conflicts with the well-being of a sentient being. These remarks, though relevant, are rather isolated in the documents; they give the impression that we need to keep a close watch on AI systems or risk losing control. Reconciling these seemingly divergent considerations does, however, seem possible: we could maintain control of AI at a certain level on an AI system while authorizing—at a lower level and within a defined framework—AI to find certain solutions to problems in a free and creative fashion.

SUGGESTED RECOMMENDATIONS

The notion of autonomy triggered more philosophical reflections than concrete recommendations, even if some recommendations from other sections are not in the report (for example ones on consent in the privacy section).

KNOWLEDGE

SUGGESTED PRINCIPLE

"AI development should foster critical thinking and protect us from propaganda and manipulation."

GENERAL OBSERVATIONS

A number of links can be made between AI and knowledge. First, from the perspective of cognitive sciences, artificial intelligence can help us understand natural intelligence, each being defined by what guides their capacity for action (SRAD). We may, then, wonder why natural intelligence should prevail over artificial intelligence because at a certain analytical level, humans and animals, just like machines, are causal systems.

In many proposals, the knowledge principle provides us with an opportunity to discuss the issues of propaganda and fake news. Seen in this light, the issue is as much about democracy as it is knowledge. We can, however, question how AI or those who produce and market it are in a position to decide what propaganda or manipulation is. It seems illegal, even dangerous to entrust them with such a responsibility. That is why the MAIEM suggests reformulating the principle to place greater emphasis on transparency: "The development of AI should not hamper critical thinking. It must also proceed in a transparent and open manner, to enable public participation in its development, scrutiny, and education."

Among other themes related to knowledge are public access to AI study results, critical thinking (MAIEM warns against echo chambers), AI education and the opacity of the algorithms, previously mentioned in the justice section. On this last point, the SRAD calls for efforts to not only improve data and algorithm transparency, but also to publish the source codes behind AI.

SUGGESTED RECOMMENDATIONS

> Measures to promote public access to the results of academic studies should be implemented. (MAIEM)

> We must encourage competition and diversity in AI applications so that they benefit society as a whole. (MAIEM)

> We must rethink the business model for social media from other news sites (MAIEM).

> All AI students and practitioners should receive advanced ethics training. (Parent)

DEMOCRACY

SUGGESTED PRINCIPLE

"AI development should foster informed participation in public life, cooperation and democratic debate."

GENERAL OBSERVATIONS

With regard to democracy, many documents (Robert, Parent, OIQ, AQT, SRAD) welcome the Declaration initiative and the opportunity it gives them to have their voice heard. MAIEM sees it as an "important contribution" to international discussions on the subject.

Others are more critical. Quintal et al. contest the very process of how the Montréal Declaration was produced. Although they are in favour of public consultation efforts, they question whether it is a way to render an existing document legitimate. More specifically, they fear that the preliminary version of the Declaration (the seven proposed principles on which this summary is built) may have strongly influenced the citizen debates: "the public should have been meaningfully engaged in deliberating the contents of the Declaration from the very beginning". For Quintal et al., this risks compromising the ultimate legitimacy of the Declaration.

These concerns are, evidently, a call for greater democracy (and transparency and critical thinking) in AI development which, to a certain extent, supports the democracy principle. Furthermore, Quintal et al. specify that democratic good will remain an empty promise if does not come with industry regulations. We also run the risk of companies using algorithms to limit the debate to issues only they deem acceptable (what we, along with the SRAD, could qualify as an epistemic issue with adverse effects on democracy). A similar argument is made by the MAIEM, which notes external regulations appear to be the best solution since, in order to protect their intellectual property, it is highly unlikely that companies will share their algorithms.

As for the principle itself, the MAIEM finds its formulation somewhat vague and deplores that it focuses on democracy when all humans do not live under this kind of regime. The MAIEM therefore suggests replacing it with a "public participation principle" which would read as follows: "The development of AI should promote the dissemination of clear and accurate information to the public to enable open and educated debate on AI and its applications, and encourage open and transparent research collaboration."

Finally, the SRAD mentions that major technology companies (such as the GAFA) hold considerable power nowadays, both political and economic—particularly because they have direct access to a tremendous amount of personal data. This can present a serious threat to democracy, as evidenced in the wake of the Cambridge Analytica affair. Furthermore, insofar as democracy demands a certain socioeconomic equality—at the risk of spiralling into an oligarchy—we must remain watchful of the growing inequalities that will automatically result from AI development. Indeed, the SRAD states that automating a task by AI comes down to transferring wealth to capital (thereby concentrating it in the hands of shareholders rather than employees replaced by AI). Unless there is a framework or regulations, AI risks amplifying the growing economic inequalities that have been observed since the 1950s.

SUGGESTED RECOMMENDATIONS

> Leading researchers in the field in our public universities must remain independent of the private sector. (Parent)

> We must break up major monopolies in the technological industry (SRAD).

> We must seriously consider the possibility of a guaranteed basic income funded by a tax on automation or on capital (SRAD).

> We must encourage new company ownership structures such as cooperatives to fight the concentration of wealth (SRAD).

< >
Montréal Declaration
Responsible AI_
</ >

PART 6

PRIORITY PROJECTS AND THEIR RECOMMENDATIONS FOR RESPONSIBLE AI DEVELOPMENT

CREDITS

TOWARDS PARTICIPATIVE GOVERNANCE OF AI

WRITTEN BY:

Nathalie Voarino, Scientific Coordinator, PhD Candidate in Bioethics, UdeM

Jean-François Gagné, Researcher at the Montreal Centre for International Studies, UdeM

CONTRIBUTIONS:

Marc-Antoine Dilhac, Associate Professor, Department of Philosophy, UdeM

Christophe Abrassart, Associate Professor in the School of Design at the Faculty of Planning, UdeM

DIGITAL LITERACY PROJECT

WRITTEN BY:

Camille Vézy, PhD Candidate in Communication Studies, UdeM

CONTRIBUTIONS:

Marie Martel, Professor in the School of Library and Information Science

Marc-Antoine Dilhac, Associate Professor, Department of Philosophy, UdeM

DIGITAL INCLUSION OF DIVERSITY PROJECT

WRITTEN BY:

Marc-Antoine Dilhac, Associate Professor, Department of Philosophy, UdeM

CONTRIBUTIONS:

Loubna Mekki-Berrada, Doctoral student in Neuropsychology, UdeM

Jihane Lamouri, Diversity Coordinator, IVADO

ENVIRONMENT PROJECT

WRITTEN BY:

Christophe Abrassart, Associate Professor in the School of Design at the Faculty of Planning, UdeM

CONTRIBUTIONS:

Alessia Zarzani, Ph.D in Planning, UdeM and Ph.D in Landscape and Environment, Université la Sapienza de Roma

Christophe Mondin, Research Professional for CIRANO

Vincent Mai, Doctoral student in Robotics, UdeM

RECOMMENDATIONS

WRITTEN BY:

Marc-Antoine Dilhac, Associate Professor, Department of Philosophy, UdeM

Christophe Abrassart, Associate Professor in the School of Design at the Faculty of Planning, UdeM

Nathalie Voarino, Scientific Coordinator, PhD Candidate in Bioethics, UdeM

CONTRIBUTIONS:

Members of the Declaration's scientific committee

TABLE OF CONTENTS

1.	**INTRODUCTION — For a creative digital transition**	**252**
2.	**TOWARDS PARTICIPATIVE GOVERNANCE OF AI**	**254**
2.1	How to Govern Algorithms: Promote Citizen Involvement	254
2.2	Not living in a world governed by algorithms: favouring human agency	258
3.	**DIGITAL LITERACY PROJECT: Ensure the lifelong development of digital skills and active citizenship**	**262**
3.1	Outfitting Canadians With Digital Skills	263
	3.1.1 The digital literacy ecosystem	264
	Outside the formal education and training system	264
	Digital literacy at school	265
	3.1.2 Professional training	266
	Developing digital skills in every sector	266
	Develop Skills Other Than Technical for AI Professionals	267
3.2	Encourage the appropriation of digital literacy by reinforcing active citizenship, diversity and solidarity	267
	3.2.1 Cyber Citizenship: Understanding, Critical Judgment and Respect	268
	Understanding, being able to act and criticize	268
	Showing Respect and Taking Responsibility	269
	Contributing to the sustainable well-being of society	269
	3.2.2 Appropriating digital culture: accessibility, inclusion and diversity	270
	Digital inclusion	270
	An Issue of Citizen Participation	270
	Inclusion Spaces: Libraries and Third Spaces	271
4.	**DIGITAL INCLUSION OF DIVERSITY PROJECT**	**272**
4.1	Algorithmic neutrality questioned	274
	Human biases and impartial machines?	274
	Discriminating Machines	275
	Biased Identity: Internet and AIS	277
4.2	Unbiasing artificial intelligence systems	280
	A Problem With Data	281
	Making Algorithms Talk	282
	Representation and Inclusiveness	285
5.	**ENVIRONMENT PROJECT: AI and environmental transition, issues and challenges for strong sustainability**	**287**
5.1	Digital transition and environmental transition: an unresolved contradiction	288
5.2	Artificial Intelligence and the Environment: Challenges and Opportunities	291
	5.2.1 Direct and indirect environmental footprint of AIS	292
	5.2.2 New predictive tools for the environmental transition	295
6.	**RECOMMENDATIONS**	**299**

FIGURES

Figure 1: Detail from the cover of Safiya Umoja Noble's book, *Algorithms of Oppression*	279
Figure 2: Search on google.com engine performed on October 29, 2018	280
Figure 3: Search performed on google.fr engine on October 29, 2018	280

1. INTRODUCTION — For a creative digital transition

The disruptive nature of digital technologies and artificial intelligence is universally recognized. But should we see the social change brought on by these technologies as an evolution, disruption or a revolution? The question is worth asking, but we will not have an answer for a few decades. What we know today is that these technologies make certain structures in our social organization obsolete and call for the creation of new structures, that they modify and reshape the work force, and that they reconfigure the urban environment, mobility and all other areas of social life.

When placed in these terms, the problem of social change necessarily recalls the "creative destruction" thesis by economist Joseph Schumpeter. The general idea is simple: a technological innovation provides economic development opportunities, and those who seize them have a decisive advantage over others. A company that develops or uses new technologies thereby becomes more efficient and can offer products that are better suited to the consumer's needs, or that satisfy new needs. The companies that refuse to yield to new technologies see their existence threatened, and even the great names eventually disappear. There are many modern-day examples: How many adults born after the year 2000 know that generations of people kept their souvenirs on photographic film that had to be developed with specialized knowledge of chemistry? Within the space of 20 years, the industry of silver gelatine photography was crushed by digital technologies, and the iconic name of Kodak is now part of the history of industrial empires. If the desire to take pictures has never been greater, it is no longer satisfied by the film industry, or only very marginally, but rather by the entire digital industry of producing and capturing images to be shared on social media.

With the rise of AI technologies, we are seeing a new phase of creative destruction, "that process of industrial mutation (...) that represents an endless revolution from within the economic structure, that constantly destroys the old and creates the new.[1]" Against the fear of AI systems (AIS) destroying jobs, of replacing workers and generating mass unemployment, certain people candidly oppose Schumpeter's thesis: Although they recognize that AIS will replace human beings in many tasks that can be automated, optimists maintain that this will create other jobs and other needs and that the job market will adjust. Society as a whole will adjust, or rather, will have to adjust:

> "This process of Creative Destruction makes up the fundamental data of capitalism: it is what capitalism, after final analysis, consists of, and every capitalist company must adapt to it, whether they like it or not.[2]"

Although Schumpeter insists on the fact that we "must adapt" to the creative destruction process, this "must" is not a moral injunction that upholds an ethical principle, but rather a pragmatic precept. If a company and a capitalist society (regardless of its political regime) wish to be sustainable, they must adapt to the realities and possibilities offered by new technologies. And yet, if adapting is necessary to brave the technological "hurricane" (the image is Schumpeter's), this hurricane will also destroy companies and organizations, it will marginalize cities and regions, and leave behind entire countries that depend on external economic activities. There can be many "losers" in this creative destruction, even if they are open to adaptation.

[1] Joseph Schumpeter (1943), *Capitalisme, socialisme et démocratie*, French transl. Gaël Fain, Paris, Payot, 1951, p. 128.

[2] Ibid.

While admitting that it is always possible to adapt—imagine that in 1995 Kodak had realized the impact that digital technology would have and had started producing the sensors now found in digital devices—such adaptation can take a lot of time for heavy structures (factories, big companies, public administrations) while technological change can happen very fast. In the case of new digital technologies and AI, change is very fast and there is no social structure capable of such change: the law, without which society becomes completely unstable, is much too slow to reform and regulate activities that legislators barely understand.

So what part will destruction play in AI development? What part will social reinvention play? How to equitably carry out a social transformation as far-reaching as the one created by the rollout of AI? Because if adapting to new AI realities is necessary, it cannot come at just any social cost, or for just any purpose. To be blunt, human beings are not very good at making predictions, and we do not know which sectors will truly be affected by the rollout of AI (self-driving vehicles, perhaps, but nothing is certain), nor if AI adaptation will be successful, or when it will occur. In the face of this uncertainty, we urgently need to find our bearings for opening up a path towards a harmonious society that integrates AI tools.

This is the crucial issue in any reflection on the digital transition. But to seriously engage in such discussions, we must not sink into pessimism, or frighten ourselves with dystopias straight out of science fiction. We will also stay clear of any naive optimism that sees in technology in general, and AI in particular, the solution to all of humanity's woes; scientist and technicist utopias have nothing to offer. Political utopias protect us from technicist naivety; they may indicate an ideal direction, but they are not rooted in the present and therefore cannot help trigger a social transformation process.

It is therefore best not to yield to utopian dreams or dystopian nightmares, but rather develop a complex realism that seriously considers the opportunities offered by technology, that does not neglect the constraints and dynamics of the present, and that tries to find action levers for guiding the implementation of AI towards the common good, social equity and human agency (autonomy).

After defining an ethical framework, we present some thoughts on how to open the way to a series of practical recommendations. This work is the result of a fruitful dialogue between experts, stakeholders and citizens. The deliberation and co-construction workshops for the Declaration had, as their explicit goal, to collectively develop concrete proposals for establishing institutional mechanisms so that AI is deployed in a socially responsible manner and respects the ethical principles of the Declaration. The deliberations helped draw up model proposals and orders of priority for the actions to be carried out over the coming months and years. Based on the results of this deliberative process, we have selected priority themes to equip public authorities, companies and citizens, and to achieve a creative digital transition of the social fabric, collective well-being, wealth and sharing: algorithmic governance; digital literacy; the inclusion of diversity; ecological sustainability.

If the world of artificial intelligence is coming tomorrow, let us keep our reasoning sharp in order to make it through the night.

2. TOWARDS PARTICIPATIVE GOVERNANCE OF AI

Governance refers to a series of formal and informal policies and procedures. It concerns both regulations and laws, standards and practices, for an organization or a series of organizations, private or public. **Algorithmic governance** refers by convention to the procedures that help guide the devices used in independent decision-making (to variable degrees) by an automated system.

However, there is a notable ambiguity attached to this term that at times refers to "how to govern artificial intelligence (AI)" and at other times to "how AI governs." This ambiguity was raised by Musiani (2013) in reference to the Governing Algorithms event which took place in New York in May 2013, and whose title could refer to either the political regulation of the technologies in question or to a certain power held by algorithms themselves to govern. This raises the question of what algorithms "can do" and to what extent they become governance artifacts through the power we bestow upon them. These two aspects are essential to the responsible management of AIS in our societies. Two main questions are therefore inherent to algorithmic governance: how will institutions manage the algorithms, and to what extent will we be living in a world governed by algorithms[1]?

2.1

HOW TO GOVERN ALGORITHMS: PROMOTING CITIZEN INVOLVEMENT

According to Antoinette Rouvroy and Thomas Berns, algorithmic governance unfolds in three steps[4]:

1. the gathering of massive quantities of data—especially by private companies;
2. the processing of this data and production of new knowledge; and
3. the use of this knowledge[5]. The issues concerning algorithmic governance are therefore inseparable from those around the data from which algorithms learn, or that they analyze. The great amount of data used enhances their effectiveness (when it comes to their training), and lends more weight to the decisions they make.

Mechanisms and proposals tied to data governance have recently been concretely implemented, as has the European Union's General Data Protection Regulation (GDPR)[6], which is not without international repercussions. Certain governments, including in Quebec, make public data accessible under various conditions[7]. The Ville de Montréal develops policies on open data[8] and open source software[9] that lean towards respect for privacy and public safety. Impact studies and risk analyses provide useful tools for decision makers[10]. Supervision mechanisms, such as the New York City

[3] Musiani, F. (2013). *Governance by algorithms. Internet Policy Review*, 2(3).

[4] For which they prefer the term "algorithmic governmentality"

[5] Rouvroy, A., & Berns, T. (2013). *Gouvernementalité algorithmique et perspectives d'émancipation.* Réseaux, (1), 163-196.

[6] China has an equivalent with "Personal Information Security Specification", whereas the United States currently prefers to not have a national policy on personal data.

[7] World Wide Web Foundation. 2008-2018. *The Open Data Barometer*: https://opendatabarometer.org

[8] http://donnees.ville.montreal.qc.ca/portail/politique-de-donnees-ouvertes/

[9] https://beta.montreal.ca/nouvelles/nouvelle-politique-au-service-de-linnovation-numerique

[10] Open Data's Impact: http://odimpact.org/; Ethics & Algorithms Toolkit: http://ethicstoolkit.ai/

Task Force for Open Data and AI, are taking shape. The Villani report in France prescribes constituting "data commons"[11]. Quebec's AI strategy raises the concept of "data trust", an idea put forward in the United Kingdom in a report entitled "Growing the artificial intelligence industry in the UK". Over forty projects around the world seek to involve civil society in reformulations of legislative frameworks[12]. Lastly, some explore techniques that allow the integration of data governance into the very design of these algorithms and insist on representativeness and genders[13].

Concerning the production of new knowledge and its uses, it is the strength and precision of the algorithmic calculations that are responsible for the new form of AIS power[14]. Processing massive amounts of data (or data mining), now possible in just a few seconds, helps establish correlations that are more or less unprecedented, but also more or less relevant. On the one hand, by relying exclusively on past data, these analysis can help inform management tools and freeze society in existing organizational paradigms (e.g. in transportation, education, justice, health care) and delay the implementation of the structural reforms that are sometimes necessary. On the other hand, the automated production of these correlations limits human intervention, and therefore the related subjectivity, giving the impression of "absolute"[5] objectivity. These issues were raised by citizens during the co-construction; they feared the dehumanizing effects of an overly "objective" approach. As Rouvroy and Berns recognize, this aspect is problematic only if these correlations are used in the framework of political and scientific interventions without ever being questioned, especially when the resulting decisions affect people.

In order to define some guidelines on the use and production of algorithmic knowledge, different proposal mechanisms have been developed. Codes of ethics have been or are in the process of being developed. The Institute of Electrical and Electronics Engineers (IEEE)[15] and the Asilomar Conference on beneficial AI are leaders in this area. Companies such as Google, Microsoft and IBM have followed suit and made public the principles they are committed to. These codes of ethics rely essentially on self-regulation tied to the growing social responsibility movement in companies. Certifications are being developed, with particular concern for prioritizing co-regulation methods, such as the International Organization for Standardization (ISO)[16] initiative. That being said, the majority of certifications are limited in scope to technical considerations and do not consider social impacts[17]. Quebec's AI strategy includes a suggestion to establish a global responsible AI organization. Impact studies are also being developed on AI use by public administrations, such as those developed by the AI NOW Institute, the Treasury Board of Canada[18], and Nesta in England. Certain states are legislating: California, for example, forces online companies to publicly disclose the use of chatbots, so that an individual can know whether he or she is dealing with a human or an AIS[19]. Algorithmic governance

[11] Cédric Villani. 2018. *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne.*

[12] See GovLab: https://crowd.law/ and https://lawmaker.io/

[13] Christian Sandvig and al. 2014. *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*; Woodrow Hartzog. 2018. *Privacy's Blueprint: The Battle to Control the Design of New Technologies.* Cambridge (MASS): Harvard University Press; Jieyu Zhao and al. 2017. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints.* https://arxiv.org/pdf/1707.09457.pdf; Tolga Bolukbasi and al., 2016. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.* https://arxiv.org/pdf/1607.06520.pdf;

[14] Cardon Dominique, *Le pouvoir des algorithmes,* Pouvoirs, 2018/1 (N° 164), p. 63-73.

[15] See the IEEE's code of ethic https://ethicsinaction.ieee.org/

[16] ISO/IEC JTC 1/SC 42: https://www.iso.org/committee/6794475.html

[17] Alessandro Mantelero. 2018. *AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment*. Computer Law & Security Review 34 (4): 754-772.

[18] Treasury Board of Canada Secretariat, *Responsible Artificial Intelligence in the Government of Canada*, Digital Disruption White Paper Series (10 April 2018) https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit

[19] Dave Gershgorn. 2018. *A California law now means chatbots have to disclose they're not human*. Quartz. October 3rd. https://qz.com/1409350/a-new-law-means-californias-bots-have-to-disclose-theyre-not-human/

can also be conceived in terms of algorithm design, including by defining objectives tied to the personal well-being, for example, by introducing demographic parity and equality in the probability of reaching AIS objectives[20].

One of the underlying issues on which participants in the the co-construction process insisted was that of shared responsibility for the management of AI development: is it up to companies or the state to develop these governance mechanisms? The influence of the companies that own the most powerful algorithms is a source of concern for many. While they decry the potential conflicts of interest, they also contest the trend toward the commoditization of data. Many are displeased with the dominant positions held by the web's giants, with sometimes unsuspected repositories of personal data held for long periods. In the background, they question the transnational data flows and, most importantly, the control exercised by Silicon Valley companies. Studies show the unexpected consequences for individuals and society, as a whole, of exploiting personal data for the purpose of maximizing profit in an oligopolistic market[21]. The power balance is asymmetrical, both between the companies themselves and between companies and individuals or society. Indeed, with respect to the companies that own massive amounts of data, some worry about monopolies forming, strengthened by mergers with smaller service suppliers[22].

But although private monopolies must be avoided, we must also beware of favouring the formation of a state monopoly on the production, ownership, access to and use of data, a monopoly which does not inspire trust among other participants in the co-construction. Some studies have found questionable practices by democratic states that have used data for surveillance purposes, and have highlighted controversial partnerships with the private sector in matters of security and defence[23]. This relationship must be clarified beyond the strategic issues, as it is being used in all of the state's areas of intervention. There should be neither private monopolies nor state monopolies: it is a diversity of players that must be maintained.

Beyond the political regime, there are differences between countries regarding algorithmic governance[24]. This raises the challenge of international cooperation and rivalries between states seeking to establish their normative hegemony[25]. The dangers of abuse of power on both sides notwithstanding, the diversity of national models for data regulation (for example those in the United States, Europe and China) cause coordination problems at the international level, but also provide opportunities for dialogue through multilateral authorities[26]. In regards to public governance, a legal and judicial framework comes with various risks and raises questions[27]: for example, by focusing too closely on the abilities of the devices at the expense of the social aspects of automation (which can undermine the protection of human values)[28]. Is it possible to regulate AI? Does the state truly have the capacity to do so?[29]

[20] David Madras, Elliot Creager, Toniann Pitassi and Richard Zemel. 2018. *Learning Adversarially Fair and Transferable Representations*. https://arxiv.org/pdf/1802.06309.pdf

[21] Frank Pasquale. 2015. The Black Box Society. *The Secret Algorithms that Control Money and Information.* Cambridge (MASS): Harvard University Press. Centre for International Governance Innovation. 2018. *Data Governance in the Digital Age. Special Report.*

[22] *Big data: Bringing competition policy to the digital era*—OECD [Internet]. [cited 2018 Sep 3]. Available from: http://www.oecd.org/competition/big-data-bringing-competition-policy-to-the-digital-era.htm

[23] Taylor Owen. 2015. *Disruptive Power. The Crisis of the State in the Digital Age.* Oxford: Oxford University Press. 168-188.

[24] Alan Dafoe. *AI Governance. A Research Agenda.* Future of Humanity Institute, University of Oxford; Bartneck, C. et al. 2006. *The influence of People's Culture and Prior Experiences with Aibo on their Attitudes towards Robots*. AI & Society: 1-14. BCG GAMMA. 2018. *Artificial Intelligence: Have no Fear the Revolution of AI at Work.* https://www.ipsos.com/en/revolution-ai-work

[25] Will Knight. 2018. *China Wants to Shape the Global Future of Artificial Intelligence*. MIT Technological Review. March 16.

[26] Susan Ariel Aaronson and Patrick Leblond. 2018. *Another Digital Divide: The Rise of Data Realms and its Implications for the WTO*. Journal of International Economic Law 21: 245-272.

[27] Scherer MU. *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*. Harvard Journal of Law & Technology, Vol. 29, No. 2, Spring 2016. http://dx.doi.org/10.2139/ssrn.2609777

[28] Ambrose ML. *Regulating the loop: ironies of automation law*. 2014; 38.

[29] Danaher J. *Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems* [Internet]. Philosophical Disquisitions. 2015 [cited 2018 Sep 3]. Available from: http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html

Sharing governance of AI development between the state and companies cannot be dissociated from a major dilemma (which emerged in the citizen discussions, regardless of the sector concerned) which opposes the protection of individual interests vs. collective interests. The answer to this dilemma is an important issue that is conditional on a normative position on which no consensus was observed during the co-construction. For example, the issues raised include the value and usefulness for the common good, or collective well-being, of sharing and pooling data (e.g. in the context of public health, crime prevention or education), versus personal privacy and the freedom to share one's data or not. Although it can be overcome, there is a fairly classic opposition between the political conception that promotes individual freedom and a non-interference space (absolute protection of data, rejection of any surveillance) with a conception that rather defends the common good, equity and process transparency, as well as policies on resource allocation and the sharing of personal information.

As for the workplace, this dilemma was basically examined from a responsibility perspective: participants identified protection of the common good according to a certain collective responsibility, arguing that it is necessary to effect a major shift towards a sharing economy and that "everyone sort of becomes their own business." Participants advocated for the individual's autonomy throughout their personal and professional lives (and the associated well-being) and expressed concern over the risk of demutualization and increased individualization in the face of social risks. Who should then be responsible for ensuring collective and individual well-being during the digital transition?

Whether it is the state or companies, the problem raised is one of the concentration of power and the verticality with which it is exercised, at the cost of a representation of civil society and a horizontal distribution of the power to organize the rollout of AI. The current context is marked by a few players who dictate the rules without, for the most part, any regard for citizens' preferences. If the discussions about governance often place public institutions and private companies in opposition, an alternative was suggested during the co-construction: participative governance, which directly reaches out to citizens by suggesting, for example, the establishment of a permanent forum for dialogue. The scientific literature shows the relevance of the contribution made by collective intelligence to technological innovation, including algorithmic governance[30]. Although the participation and collaboration of stakeholders take time, they are still valuable[31]. The organization of "hybrid forums" where citizens, experts and administrations collaborate around complex objects like AIS is especially justifiable in an uncertain world where at any moment sociotechnical controversies can erupt, in which no player can claim omniscience[32]. Some have therefore tried to open the algorithms to the public[33]. However, the perceptions, preferences and interests of citizens remain, in the vast majority of cases, too small a concern in the decision making around a responsible rollout of AI.

In the optics of this participative governance, citizens highlighted the importance of user contributions to the design and management of AI tools. This participation could take the form of a collective experimentation based on user experience (design thinking) through open-source prototypes. This material, accessible to all, constitutes a digital common good (for example, open source software or data commons[34]), which seems characteristic of the digital rollout as it currently stands. "The digital rollout is characterized by the creation of public goods by Internet communities. This process supposes the emergence of significantly new

[30] Geoff Mulgan. 2017. *Big Mind: How Collective Intelligence Can Change Our World.* Princeton: Princeton University Press. Danaher et al. 2017. *Algorithmic Governance: Developing a Research Agenda through the Power of Collective Intelligence*. Big Data & Society: 1-27

[31] Elizabeth F. Cohen. *The Political Value of Time*. Cambridge: Cambridge University Press

[32] Michel Callon, Pierre Lascoumes, et Yannick Barthe, 2001, *Agir dans un monde incertain. Essai sur la démocratie technique*, Paris, Le Seuil, "La couleur des idées".

[33] See https://algoritmi.pybossa.com

[34] The Villani report recommends establishing "data commons", which would encourage economic players to pool their data and would give public stakeholders more weight.

organizational structures supported by information technologies, especially open source movements and the Web 2.0." [translation][35] More than a simple form of ownership, it is a cooperative organizational model that guarantees horizontal exchanges between peers, as well as freedom of expression[36]. This organization relies on methods of regulation agreed upon by the actors themselves[27]. This type of governance is not without its own set of challenges, particularly vulnerable to different forms of enclosure (a reduction in shared uses) by both the state and companies[37]. At a later step, we must envision that the social parameters of algorithms will be subjected to citizen deliberation, or even better: citizen coding. This coding should not involve skills superior to those guaranteed through the acquisition of digital literacy, as we will see in the next section, and will not require consulting the entire population, but rather multiple deliberative groups.

Regardless of the actor, the participants insisted that there is a collective responsibility for the social impacts of AI. Behind this idea lies a concern, however: the speed at which technology is changing leaves little time for citizen deliberation and political reflection. To meet these different challenges, it seemed relevant to promote a form of governance that relies on citizen involvement, including to guarantee that the AI rollout reflects society's fundamental principles and values. It therefore appears essential to create inclusive means of consultations that involve citizens in all their diversity, at different steps in the oversight process for AI responsible development (see Section 6 of this report, Recommendation 1). This collective participation should take place for AI design, as well as to provide oversight based on user feedback on problems as they arise.

2.2

NOT LIVING IN A WORLD GOVERNED BY ALGORITHMS: FAVOURING HUMAN AGENCY

Citizens who took part in the co-construction activities support the idea of a certain "digital humanism". This implies that AIS integrate fundamental ethical principles or human values in order to protect everyone's interests, including the right to privacy, protection of the environment, even the preservation of what defines us as human beings. They fear a dehumanization of the various sectors of activity affected by AI development, by reducing human beings to quantifiable data. They also worry that AI expertise will be valued over human expertise, and that it will become difficult to maintain control over the algorithms and their decisions. These concerns refer to the second conception of algorithmic governance, i.e. "how AI governs us".

Algorithms already impact our daily lives. Different authors signal the widespread use of various computational methods, necessarily approximative and standardized, to evaluate individuals, as well as their potentially adverse and unforeseen consequences[38]. Here the danger lies in the omnipotence of the computer language that shapes this world of possibilities, with no concern for the inherent subtleties of social context[39]. The use of marketing algorithms that recommend products based on your purchase history and products consulted is one example of the appearance of algorithms that "govern" by guiding the choices of consumers[40]. The "digital profiles" are therefore used, sometimes unbeknownst to the individuals concerned, for different purposes, at the risk of

[35] Ruzé E. *La constitution et la gouvernance des biens communs numériques ancillaires dans les communautés de l'Internet. Le cas du wiki de la communauté open source WordPress*. Management & Avenir. 2013;(65):189–205.

[36] Crosnier HL. *Communs numériques et communs de la connaissance. Introduction*. tic&société. 2018 May 31;(Vol. 12, N° 1):1–12.

[37] Crosnier HL. *Une bonne nouvelle pour la théorie des biens communs*. Vacarme. 2011;(56):92–4.

[38] Jerry Z. Muller. 2018. *Tyranny of the Metrics*. New Jersey: Oxford University Press; Andrea Saltelli and Mario Giampietro. 2017. *What Is Wrong with Evidence Based Policy, and How Can it Be Improved?* Futures 91:62-71. Joshua Newman. 2016. *Deconstructing the Debate over Evidence-Based Policy*. Critical Policy Studies 11 (2): 211-226.

[39] Tarleton Gillepsie. 2012. *The Relevance of Algorithms*. Tarleton Gillespie, Pablo Bocskowski and Kristen Foot (dir.). *Media Technologies.* Cambridge (MA): Cambridge University Press; Ed. Finn, 2017. *What Algorithms Want—Imagination in the Age of Computing,* Cambridge (MA): MIT Press.

[40] Ibekwe-Sanjuan, Fidelia. *Big Data, Big machines, Big Science: vers une société sans sujet et sans causalité?. XIXe Congrès de la Sfsic. Penser les techniques et les technologies: Apports des Sciences de l'Information et de la Communication et perspectives de recherches*. 2014.

replacing their true identities[28]. Therefore: "Leaving digital traces becomes synonymous with normalcy, but at the price of permanently exposing oneself. Not to leave a digital trace becomes suspicious contrarian activity and can trigger increased surveillance. It is therefore no longer possible to escape being circled by electronic devices."[28] The risk then becomes that an individual can be placed in danger through desubjectivation[41]. The citizens argued, however, that a person's situation should not be reduced to quantifiable factors.

In order to prevent a situation where algorithms "govern" us, it appears necessary, on the one hand, to temper the power we grant them and, on the other hand, to foster AIS development that promotes **human agency**, i.e. the individual's ability to act. [42]Indeed, considering the increasingly autonomous nature of AI, some philosophers have reconsidered the concept of "moral agency" that had until now only been attributed to human beings[43]. This means that by "making decisions," algorithms would bear a kind of responsibility towards the consequences of the actions resulting from their recommendations, thereby becoming "agents" or actors in society. The automation of data analysis and decisions made by AIS raise important questions regarding sharing control between humans and algorithms[44], in particular because it is not yet possible to explain to users the path that an AIS has taken to make a decision (the famous AI black box). There are concerns regarding the rollout of algorithms and their negative impact on free will and individual autonomy[45], which could potentially impair the ability of individuals to assume certain responsibilities (thereby impairing their agency). The citizens raised the issue of a risk that, by giving AI too much power or sovereignty in decision making, humans would be disempowered or lose skills. Some have even claimed that agency deserves its own principle in the Montréal Declaration (see Part 7, Results of Winter Co-construction).

However, it is important to highlight that the algorithms' calculation rules are procedural and not substantive, meaning that the algorithms have no real understanding of the information they handle, or even the results they produce[37]. Therefore, it is the human beings behind their programming, those who implement AIS in their organizations, or those who use their recommendations, who must be held responsible for the consequences of the actions and decisions made by AIS. In other words, humans are the only agents of algorithmic governance; they are the ones who must make the final decisions and be accountable for the adverse consequences—and benefits—of AIS use.

But here is cause for doubt: if AIS do not govern in the human sense of the term, it is entirely possible that they are the agents of a governance by procedure, and not by reflecting on the social and ethical substance of the decisions they are making. That is why we must normatively claim, as established by the participants of the Montréal Declaration, that final decisions must be submitted to human control, namely for the moral, functional and political aspects of AI, despite (and against) its procedural efficiency. This recommendation aligns with many other international reports, such as that from CNIL in France with the unequivocal title: "Comment permettre à l'homme de garder la main?" (How can man keep the upper hand?)[46]. A minority considers it acceptable to delegate microdecisions to algorithms, depending on the gravity of the consequences and the complexity of the phenomenon. This position is in line with that of the participants who insist on the need to keep a human in the loop of algorithmic decisions[47], which is all the more important when it comes to decisions

[41] Rouvroy, A., & Berns, T. (2013). *Gouvernementalité algorithmique et perspectives d'émancipation*. Réseaux, (1), 163-196.

[42] More specifically, agency can refer to the ability humans have to think about what they value, set goals and achieve them (Isle Oosterlaken, *Technology and human development*, Routledge, 2015, p. 5).

[43] Noorman M. Computing and Moral Responsibility. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy* [Internet]. Winter 2016. Metaphysics Research Lab, Stanford University; 2016 [cited 2017 Jun 8]. Available from: https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/

[44] Musiani, F. (2013). *Governance by algorithms.* Internet Policy Review, 2(3).

[45] Cardon Dominique, *Le pouvoir des algorithmes*, Pouvoirs, 2018/1 (N° 164), p. 63-73.

[46] CNIL, report *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*, 2017.

[47] Some even suggest a model that would include different stakeholders in the decision-making process based on the parameters of a social contract (society in the loop). See: Rahwan, Iyad. *Society-in-the-loop: programming the algorithmic social contract.* Ethics and Information Technology 20.1 (2018): 5-14.

with serious consequences (such as the decision to kill[48]).

Both in the short and mid-term, humans appear destined to keep control over AI[49]. Exercising their agency supposes both preserving certain skills and ensuring access to knowledge (for more information, see the section on digital literacy). In other words, this involves establishing governance that allows access to the skills and knowledge required not only for individuals to exercise their agency, but also the governance of organizations that roll out AI and must maintain a reflective, critical and learning relationship with these tools.

One of the manifestations of this exercise in terms of governance is obtaining free and informed consent from the people who use AIS or are subjected to its analysis. In this perspective, the citizens argued that it is absolutely necessary for an individual to know who is using their data and the intentions of the acquirer, in order to guarantee informed consent. Other citizens felt that an individual should have access to an understandable justification. Knowing the margin of error of the option indicated by an algorithm, and the objectives guiding its recommendations, also appeared crucial to the citizens involved in the co-construction. This transparency requirement is not only a necessary condition for trust, but a key element in exercising agency. In this sense, the citizens believe that organizations should assume their responsibilities and take appropriate measures so that the "burden of consent" does not rest solely on the user's shoulders.

However, much has been written by legal experts about the concept of "informed" consent: it is being received in conditions that are further and further away from the spirit of law[50]. Even more problematic for urban planners is the acquisition of data without explicit consent, namely in the public space with smart cities and connected objects[51]. As for the health care sector, other actors question whether it is possible, under current conditions, to obtain truly informed consent from patients given the uses that are being made made of AI, in particular in regards to the protection of privacy and confidentiality, which are threatened by the exponential reuse of biomedical data[52]. It does now seem difficult to foresee, *a priori*, all the potential uses of every set of data produced, and therefore warn individuals. In this context, it becomes imperative to revisit the concept of privacy beyond the legal corpus[53]. Certain philosophers introduce the idea of a right to interiority[54], while programmers experiment, to mixed results[55], with personal data de-identification techniques to prevent (re)identification.

For many researchers, the opacity of neural networks is precisely the core of the problem[56]. And in the public sector, this is a major issue, as algorithms are making decisions that have a major impact on daily life[57]. Without any explanation, especially in the case of mistakes and malfunctions, and without any recourse, the prejudices committed may unjustly penalize individuals[58], especially since there are often no feedback mechanisms to address the imperfections of automated systems, since the calculations remain cryptic and the statistics,

[48] Peter Asaro. 2012. *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making*. International Review of the Red Cross 94 (886): 687-709.

[49] AI Timeline Surveys: https://aiimpacts.org/ai-timeline-surveys/

[50] Fred H. Cate and Viktor Mayer-Schönberger. 2013. "Notice and Consent in a World of Big Data". International Data Privacy Law 3 (2): 67-73. Omer Tene and Jules Polonetsky. 2013. *Big Data for All: Privacy and User Control in the Age of Analytics.* Northwestern Journal of Technology and Intellectual Property 11 (5): 239-272.

[51] Rob Kitchin. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* Thousand Oak (CA): Sage.

[52] Mittelstadt BD, Floridi L. *The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts*. Sci Eng Ethics. 2016 Apr; 22(2):303–41.

[53] Colin J. Bennett and Charles Raab. 2018. *Revisiting the Governance of Privacy: Contemporary Policy Instruments in Global Perspective*. Regulation & Governance: 1-18; Neil M. Richards and Jonathan H. King. 2014. *Big Data Ethics*, Wake Forest Law Review 49:393-432.

[54] Sara Champagne. 2018. *Trois questions sur la vie privée au philosophe Jocelyn Maclure*. Le Devoir. March 17.

[55] *Article 29 work group on data protection. Avis 05/2014 sur les Techniques d'anonymisation*. https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr_0.pdf

[56] Mike Ananny and Kate Crawford. 2018. *Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability.* New Media & Society 20 (3): 973-989.

[57] Cathy O'Neil. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threaten Democracy.* New York: Broadway Book.

[58] ProPublica. *Machine Bias*. https://www.propublica.org/series/machine-bias; Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.*, New York: St. Martin's Press; Mittelstadt et al. 2016. *The Ethics of Algorithms: Mapping the Debate*. Big Data & Society: 1-21.

hidden[59]. It is therefore for control purposes that this transparency is required, in particular to ensure human responsibility for abuse (and thereby limit it). For example, certain studies describe the discrimination generated by the many biases inherent to AIS. One of them employs epistemological considerations related to scientific objectivity: data is a social construct, a value judgment, it is not neutral[60]. Although the problem of data reliability is well documented in the history of science, the risk of bias takes on alarming proportions with AI due to its scale: every individual is a potential victim, even if not everyone will be affected[61] (for more information, see the section on digital inclusion of diversity).

In this respect, it appears essential to promote and ensure that AIS are developed in such a way so as to preserve and even increase the abilities of people and organizations. This aspect echoes the FACIL Declaration, which advocates digital technologies derived from knowledge that is developed collectively and advocates for the protection of citizen's abilities[62]. Along the same lines, it is important to mention ATM (the appropriate technology movement), which is based on the capabilities approach[63] to reflect technological development. According to this movement, there is no reason to assume that the most advanced technology is necessarily the best option; the real issue is the true value added by technological developments in terms of human capabilities. Two aspects of the capabilities approach are particularly relevant here. First, it involves concentrating on individuals' abilities and functioning rather than only on the means (like resources, for example). Second, it involves paying special attention to human diversity. Respect for this diversity is one of the main reasons for focusing development objectives on expanding human capabilities rather on access to resources.

Achieving well-being is the main demand under this approach. Agency is one of its key concepts; it assumes that individuals are not passive receptors but rather active participants in development (in this case, technological development). Following this train of thought, communities must guide technological development (which is aligned with participative governance) so that it reflects their values and objectives.

So in the interests of promoting the implementation of adapted governance, we saw a need to delve deeper into three priority areas of intervention in order to formulate recommendations on public policies. These areas are:

1. a project on digital literacy issues (to ensure the development of everyone's digital skills);
2. a project on the issues related to the inclusion of diversity; and
3. a project on the environment (to guarantee sustainable well-being and strong ecological sustainability in AIS development).

These three projects emphasize the essential (though not exhaustive) conditions for establishing a governance that seeks to underpin the well-being of individuals in all their diversity and promote their agency, including in the context of participative governance. We consider these conditions essential to ensuring that algorithms have a positive impact on the lives of individuals, and that everyone can be an actor in his or her digital reality, with an eye toward collective responsibility.

[59] Cathy O'Neil. Op. cit.

[60] Alex Campolo et al. 2017. AI NOW Report: 15. Luciano Floridi and Mariarosaria Taddeo. 2016. *What is Data Ethics*. Philosophical Transactions of the Royal Society 374:1-5; Erna Ruijer et al. 2018. *Open Data Work: Understanding Open Data Usage from a Practice Lens*. International Review of Administrative Sciences 0 (0): 1-17.

[61] Cathy O'Neil. Op. cit.

[62] FACIL Digital Commons Declaration: https://wiki.facil.qc.ca/view/D%C3%A9claration_des_communs_num%C3%A9riques

[63] The capabilities approach is derived from the work of Amartya Sen and Martha Nussbaum."These two thinkers both argue that assessment of development progress should not be made on terms of income or resource possession, but in terms of valuable individual human capabilities – or what people are effectively able to do and be". (Isle Oosterlaken, *Technology and human development*, Routledge, 2015., p. 2). Therefore, a capability can be understood to be the ability to carry out a fundamental human good such as traveling, staying healthy or developing one's mind.

[64] Isle Oosterlaken, *Technology and human development*, Routledge, 2015.

3. DIGITAL LITERACY PROJECT: Ensuring lifelong development of digital skills and active citizenship

Montréal Declaration for Responsible AI, Principle 2.4:

"It is crucial to empower citizens regarding digital technologies by ensuring access to different types of knowledge, the development of structuring skills (digital and media literacy), and the rise of critical thinking."

Digital literacy is recognized by organizations such as UNESCO and the OECD as being **central to social and citizen involvement in an information society and knowledge economy.** It is defined as "the ability to define, access, manage, integrate, communicate, evaluate and create information safely and appropriately through digital technologies and networked devices for participation in economic and social life."[65] It includes skills that are variously referred to as computer literacy, ICT literacy, information literacy, data literacy and media literacy[66]. Digital literacy is therefore not limited solely to knowing how to use digital tools, it also includes a critical dimension that leads to knowing how to make informed decisions regarding this use.

In an information society that rests, above all else, on a civilization of the written word, digital literacy relies on the ability to understand and use written information in everyday life (functional literacy). It is therefore part of a continuum running from basic literacy to the ability to understand and interact with AIS in informed manner.

LITERACY — DIGITAL LITERACY — AI LITERACY

[65] UNESCO (March 2018). *A draft report on a global framework on digital literacy skills for indicator 4.4.2: Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills.* http://gaml.cite.hku.hk/wp-content/uploads/2018/03/DLGF-draft-report-for-online-consultation-all-gaml.pdf p. 3.

[66] Ibid.

During the Montréal Declaration's citizen deliberations, the digital literacy issue was discussed in every field. Citizens highlighted the **need to educate the population about the issues and practices** in artificial intelligence. This training would provide **both the technical and critical skills** required for any individual to act in an independent, informed and responsible manner as a **worker and citizen** in a society in transition. The main goals are therefore to foster the **development of a good understanding and critical thinking** about how artificial intelligence systems (AIS) operate, their use and the related new standards, in particular regarding personal data. Digital literacy has therefore become essential to citizens as a set of skills to maintain, in particular, **collective vigilance in order to develop and use AIS in a responsible manner.**

Although young people are targeted by digital literacy as early as grade school, it is also for students, regardless of their specialization, as well as professionals in every field (especially health care, education, justice, human resources and public administration). AIS designers and programmers are also concerned by digital literacy, in particular because of the need to "integrate training on ethics related to AI issues and technologies into the engineering curriculum and in continuing education" (Ordre des ingénieurs du Québec brief, Recommendation 5).

To this end, the main potential solutions suggested during the Declaration's co-construction process were to develop digital literacy at every age, through both technical education and training in ethics. This education would be dispensed through formal channels such as schools, universities and ongoing professional development, but also through "public training" in AI (see Part 3, Report on the Winter Co-construction Workshop Results, section 5.2) and the related digital realities in order to reach the entire Canadian population.

Furthermore, citizens raised two social justice issues regarding digital literacy: it must be developed in an accessible manner for all, across all of Canada, and must also be developed so as to maintain a diversity of learning profiles and paying attention to the various types of intelligence. This requires developing solutions so that digital literacy training is structurally accessible and inclusive and both promotes and reflects diversity.

Given these ideas generated by the citizens' deliberations, we will explore digital literacy development in two stages in order to present recommendations aligned with the Montréal Declaration principles, i.e. autonomy, responsibility, equity, diversity and solidarity. The main objective is to ensure the development of digital skills throughout one's lifetime, whether through formal channels (school, university, professional training) or through informal channels (outside of these systems). This digital literacy development as lifelong learning has its own two objectives:

1. to develop the human capital of Canadians by equipping them with digital skills; and
2. to encourage the appropriation of digital literacy by reinforcing active citizenship, diversity and collaboration between the members of a community, thereby fostering the development of a learning society.

3.1

EQUIPPING CANADIANS WITH DIGITAL SKILLS

Digital skills are the ability to find, understand, organize, evaluate, create and disseminate information through digital technologies; they allow us to reach objectives related to learning, work and social participation. The reinforcement of digital skills represents an innovation and economic development issue across Canada which aims to develop the skills of Canadians to give them easier access to high-paying jobs and to grow the middle class, as set forth in the *Innovation and Skills Plan*[67]. The human capital approach[68] therefore seems to be well suited to this purpose: it involves investing in the skills and

[67] Canada. Department of Finance. (2017). *Building a Strong Middel Class. Chapiter 1: Skills, Innovation and Middle Class Jobs*. pp. 47-85. Ottawa: Department of Finance. Viewed online at https://www.budget.gc.ca/2017/docs/plan/budget-2017-fr.pdf pp. 48-52.

[68] Schultz, T. W. (1961). *Investment in human capital.* The American Economic Review, 51(1), 1–17.; Becker, G. S. (1975). *Human capital: A theoretical and empirical analysis with special reference to education*. Chicago, IL: University of Chicago Press.

knowledge that individuals can acquire to foster economic growth and international competitiveness by training a competent workforce. This takes the form of, among other things, investments made by Innovation, Science and Economic Development of Canada (ISED) to develop digital literacy initiatives, but also by the artificial intelligence pan-Canadian strategy led by the Canadian Institute for Advanced Research (CIFAR), as well as national workforce strategies such as the one put forward by Québec's Ministère du Travail, de l'Emploi et de la Solidarité sociale (TESS) in support of the digital transition.

In the context of a society in transition, digital literacy first presents itself in terms of the skills it helps workers acquire to gain access to jobs and/or ensure the transformation of existing jobs. To this end, measures guaranteeing equal access to the development of these skills and equal opportunities to gain access to these jobs should be put forward.

These digital skills can be divided into three types, combining technological knowledge and critical judgment[67]:

1. Basic digital skills, which every individual needs in order to take part in modern society. This could include how to find reliable information (media or information literacy), communicating with other individuals in a considerate and safe fashion, learning to use data (data literacy), and using different types of software and apps to confidently interact with technology.
2. Skills pertaining to a specific work sector whose jobs will be transformed, requiring more interaction with AIS so workers will need to use them in a responsible manner.
3. The skills of digital professionals, representing the set of skills required to develop new technologies, services and products. This includes, for example, mastering various programming languages, data analysis methods and automatic learning techniques.

In a lifelong learning perspective, these skills will need to be developed both in the formal systems of schools, universities and professional training, but also increasingly outside of these systems, through initiatives led by private companies and not-for-profit organizations. A balance needs to be struck to encourage links between educational technology companies, not-for-profits, schools and universities, so that digital education is developed as a public asset accessible to all.

3.1.1 The digital literacy ecosystem

OUTSIDE THE FORMAL EDUCATION AND TRAINING SYSTEM

Canada already has many education and training programs for developing digital literacy. **Many organizations outside the formal education system** are developing and offering a wide range of activities.

Innovation, Science and Economic Development Canada (ISED) launched **two major programs to develop digital literacy initiatives: CanCode** ($50 million invested over a two-year period, starting in 2017-2018) and the **Digital Literacy Exchange Program (DLEP)** ($29.5 million invested from 2018 to 2022).

The initiatives funded by CanCode encourage educational opportunities for **coding** and **digital skills** development for Canadian youth from kindergarten to grade 12 (K-12)[70]. The program also funds the training and professional development of new teachers through MediaSmarts, which creates many online resources[71]. The DLEP funds projects aimed at a larger audience in order to "equip Canadians with the necessary skills to engage with computers, mobile devices and the Internet safely, securely and effectively"[72].

The approaches used by organizations **outside the formal education** system are diverse—mentorship,

[69] From Huynh, A., Lo, M., & Vu, V. (2018). *Levelling Up: The Quest for Digital Literacy*. Toronto: Brookfield Institute for Innovation + Entrepreneurship. p. 4-5. Viewed online at http://www.deslibris.ca/ID/10097218

[70] For an overview of initiatives financed by the CanCode program: https://www.ic.gc.ca/eic/site/121.nsf/fra/00003.html

[71] http://habilomedias.ca/ressources-pédagogiques

[72] Government of Canada. Innovation, Science and Economic Development. (2018) *Digital Literacy Exchange Program*. Ottawa: Innovation, Science and Economic Development Canada. Viewed online at http://www.ic.gc.ca/eic/site/102.nsf/fra/accueil

paid training, programs in community centres, workshops in libraries, online courses—and are intended for many audiences, from youth to seniors, including post-secondary students and professionals. The activities consist of intensive training (bootcamps) to learn different programming languages (e.g. Lighthouse Labs, Canada Learning Code), techno-creative workshops in fab labs (Communautique) and libraries (TechnoCultureClub) to learn 3D printing, for example, mobile application creation competitions to encourage technological entrepreneurship among young girls (Technovation Montréal), online resources on digital literacy for parents, children and teachers (MediaSmarts), and many others[73]. The development of online courses also helps validate knowledge or simply independently nurture curiosity. Many of these initiatives are funded through federal or provincial subsidies (such as CanCode and DLEP), but also through private investments. Such is the case for Ubisoft, for example, which invests over $8 million in the CODEX program, which brings together "a group of initiatives targeting all levels of education where the video game is a source of motivation and a learning engine toward the development of Quebec's future techno-creative generations"[74].

Although **the offer of training and educational activities outside the formal system** is rich and diversified, **it is not clearly organized and it can be difficult to find** the one best suited to one's needs based on age, knowledge level and interests. It is, however, worth mentioning the existence of a few tools that help guide people, either through online mentorship (Academos) or by listing activities that develop digital skills (Ma Vie Techno).

A better structuring of this ecosystem benefits individuals looking for digital training at any age, as well as actors in the community (start-ups, small or mid-sized companies, not-for-profits, community centres, etc.) that could further share their practices, but also decision makers whose choices could be made easier by having a better overview of the needs and realities of the actors that are taking part in establishing tomorrow's schools and universities and making lifelong learning possible[75].

DIGITAL LITERACY AT SCHOOL

Digital education is dispensed more and more through **formal channels,** at the elementary and high school level, as well as post-secondary institutions, through new programs and the implementation of technology as a learning tool.

In Quebec, digital literacy does not yet appear in the *Programme de formation de l'école québécoise*. It is, however, similar to media studies, which represent a general training field (like health, entrepreneurship, citizenship and the environment), but it does not represent a discipline like French, mathematics, art or history and geography[76]. The *Plan d'action numérique en éducation et en enseignement supérieur of the Ministère de l'Éducation et de l'Enseignement supérieur*[77] (MEES) does, however, introduce 3 guidelines (and 33 measures) intended to support the development of digital education:

Guideline 1: Support the development of digital skills among youth and adults

Guideline 2: Capitalize on digital technologies as a driver of added value in teaching and learning practices.

Guideline 3: Create an environment conducive to a digital rollout throughout the entire education system.

[73] The *Brookfield Institute for Innovation + Entrepreneurship* report (see note 5) offers a rich overview of the organizations and types of activities offered on Canadian soil.

[74] https://montreal.ubisoft.com/fr/programme-codex/

[75] This could be inspired by the EdTech observatory in France, which brings together digital players for education and training: http://www.observatoire-edtech.com

[76] HabiloMédias. (2016). *Québec — Aperçu de l'éducation aux médias*. Viewed online at http://habilomedias.ca/ressources-pedagogiques/resultats-dapprentissage-en-education-aux-medias-et-litteratie-numerique-par-province-et-territoire/quebec-apercu-de-leducation-aux-medias

[77] Québec. MEES. (2018). *Plan d'action numérique en éducation et enseignement supérieur.* Quebec: Ministère de l'Éducation et de l'Enseignement supérieur. Viewed online at http://www.education.gouv.qc.ca/fileadmin/site_web/documents/ministere/PAN_Plan_action_VF.pdf

However, this writing digital literacy training is dispensed randomly, without evaluation, at the initiative of teachers and principals, whether at the elementary, high school, college or university level. There are many initiatives to structure digital skills training, whether for students or teachers and professors. Such is the case with REPTIC[78], for example, which develops activities and establishes a profile of information, cognitive, methodological and technological skills, or the Association of College & Research Libraries (ACRL), which created a model for information literacy in higher learning[79]. These kinds of initiatives would benefit from being clearly integrated into education policy in order to have a greater impact and help structure digital literacy training.

3.1.2 Professional training

DEVELOPING DIGITAL SKILLS IN EVERY SECTOR

In terms of professional training, the development of digital skills is put forward, in particular in the *National Workforce Strategy* 2018-2023[80] from Quebec's ministère du Travail, Emploi et Solidarité sociale (TESS), in order to increase productivity in the workforce through ongoing training[81]. The strategy targets every worker, whether he or she holds a job or not.

Jobless individuals will be able to reach out to Services Québec, to training establishments, to organizations specializing in employability development and to training companies that will "collaborate to identify training and learning needs, expand training offers, integrate digital technology skills into job search assistance and properly prepare the workforce to acquire digital technology skills."[82] People who already hold a job requiring them to develop or upgrade their digital skills could reach out to Emploi Québec, which will "increase its purchases of part-time training based on the needs defined in the regions of Quebec"[83]. Upgrading workers' digital skills is therefore a part of the TESS strategy, but it is worth noting that the strategy does not mention the need for workers to adapt to the growing number of AIS and automated systems, which will transform many occupations.

Ongoing training must also be offered and coordinated by employers, especially when their employees' jobs are being transformed by the use of AI for different tasks, as it is the case in health care, education, justice and public and private administrations. Such training should then not only **allow workers to acquire the technical skills to know how to use AIS in day-to-day tasks,** but it must also **encourage these professionals using AIS to do so responsibly** by making them aware of the ethical and social dimensions of this use. This training could focus on making decisions with AIS assistance so that human intervention is not excluded (see the responsibility principle)—especially when the decision affects a person's life, quality of life or reputation—and so that the measure of the decision's social and ethical implications is always taken into consideration and becomes a professional reflex.

To this end, **codes of ethics** (see Part 4, Report on the Results of the Winter Co-construction Workshops, section 5.2) or a form of **"permit to use AI and algorithms"**[84] in specific sectors (health care, marketing, human resources, justice, education, public administration) could be created and obtained **after completing specific training modules offered by universities and specialized schools.** Every professional interacting with AIS decision assistance tools should also receive **appropriate training allowing them to make responsible use of these tools and be able to justify their decisions** (see the democratic participation principle).

[78] https://www.reptic.qc.ca/

[79] English version: http://www.ala.org/acrl/standards/ilframework ; French version: http://ptc.uquebec.ca/pdci/referentiel-de-competences-informationnelles-en-enseignement-superieur

[80] Québec. TESS. (2018). *National Workforce Strategy 2018-2023. Quebec in the Full Employment Era*. Québec: Ministère Travail, Emploi et Solidarité sociale. Viewed online at https://www.mtess.gouv.qc.ca/publications/pdf/Strat-nationale_mo.PDF

[81] Title of axis 3.3 of the *National Workforce Strategy 2018-2023*

[82] Measure 41 of the *National Workforce Strategy 2018-2023*, p. 70

[83] Ibid.

[84] p. 55. CNIL. (2017). *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*. CNIL. Viewed online at https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

DEVELOP NON-TECHNICAL SKILLS IN AI PROFESSIONALS

AI skills training has received considerable higher education funding, in particular through the Canadian Institute for Advanced Research (CIFAR). This organization is tasked with operationalizing the **pan-Canadian artificial intelligence strategy**, which aims to maintain and develop research excellence in Canada[85] through four major goals:

1. to increase the number of outstanding artificial intelligence researchers and skilled graduates in Canada;
2. to establish interconnected nodes of scientific excellence in Canada's three major centres for artificial intelligence in Edmonton, Montreal and Toronto;
3. to develop global thought leadership on the economic, ethical, policy and legal implications of advances in artificial intelligence; and
4. to support a national research community on artificial intelligence[16].

Over half of its budget ($86.5 million) is devoted to creating artificial intelligence research chairs to attract and retain the best university researchers in the fields of deep learning and learning through reinforcement. While these chairs seem to be exclusively tied to the computing world, an AI and Society program has also been announced to fund groups working on the political and economic implications of artificial intelligence in order to inform politicians and the general public about these issues.

Funding the creation of knowledge on AI therefore includes ethical, political, economic and social reflection on AI. This reflection should be transmitted to students and AI researchers so they can integrate these issues into their AI development practices. Initiatives are emerging in this respect, such as the responsible computing challenge initiated by the Mozilla foundation to explore new ways to teach ethics to computer science students[86]. Better trained on the social and ethical issues surrounding the AIS and data acquisition and archiving systems (DAAS) they create or use, and made aware of their share of responsibility in the development of such systems, designers and programmers could choose to use, or not use, certain AI algorithms and devices once they know more about their potential effects[87].

3.2

ENCOURAGE THE APPROPRIATION OF DIGITAL LITERACY BY REINFORCING ACTIVE CITIZENSHIP, DIVERSITY AND SOLIDARITY

The lifelong training in digital skills, whether they are basic skills or professional skills, thus requires developing technical learning and raising awareness for informed use and socially responsible conduct. Digital literacy therefore includes data literacy, media literacy and an artificial intelligence literacy that includes the analysis and critical evaluation of AIS issues. It is not only an issue of economic development achieved by reinforcing each individual's human capital, but also an educational and humanist issue[88] which aims to promote active citizenship in the digital space.

By integrating digital literacy through a lifelong learning (LLL) dynamic, we highlight the humanist and democratic values of inclusion and emancipation on which LLL relies, according to UNESCO:

> "The role of lifelong learning is critical in addressing global educational issues and challenges.

[85] CIFAR. (2017). *Pan-Canadian Artificial Intelligence Strategy Overview* [CIFAR]. Viewed online on June 23, 2018, at https://www.cifar.ca/assets/pan-canadian-artificial-intelligence-strategy-overview/

[86] https://foundation.mozilla.org/en/initiatives/responsible-cs/ ; https://www.fastcompany.com/90248074/mozillas-ambitious-plan-to-teach-ethics-in-the-age-of-evil-tech

[87] See Part 4, *Overview of international recommendations for AI ethics* (report from the Royal Society) + Part 5, *Report of online coconstruction and submissions received* (OIQ + AI Ethics meetup and survey answers)

[88] Along the lines of Regmi, Kapi Dev. (2015). Lifelong learning: *Foundational models, underlying assumptions and critiques.* In International Review of Education, 61:133-151.

Lifelong learning "from cradle to grave" is a philosophy, a conceptual framework and an organising principle of all forms of education, based on inclusive, emancipatory, humanistic and democratic values; it is all-encompassing and integral to the vision of a knowledge-based society"[89].

Digital literacy is therefore part of the knowledge which allows each person to acquire the knowledge and skills required to realize his or her aspirations and contribute to a society[90] in which digital technologies play an ever-growing part. Understood as a personal and collective growth issue, it must be developed in an accessible and inclusive manner, reinforcing the solidarity of active citizens in a learning society. In the face of a discourse that promotes the development of digital skills in the name of an employability imperative, digital literacy should develop in a way that favours a diversity of intelligences, profiles, genders and generations, in order to slow down a certain standardization of society by maintaining its diversity.

3.2.1 Cyber Citizenship: Understanding, Critical Judgment and Respect

The concept of "cyber citizenship" refers to the exercising of one's fundamental rights, political skills (such as participating in debates and public decisions), and civility obligations in a digital environment. Cyber citizens develop or use digital tools to participate in political life. They can also define themselves as members of a digital community that takes political action.

This concept raises five major issues: freedom of expression and quality of information, the individual and social responsibility of digital actors, transparency, respect of privacy, and justice.[91]

UNDERSTANDING AND BEING ABLE TO ACT AND CRITICIZE

Cyber citizenship relies on the principles of respect for autonomy, responsibility, but also democratic participation and protection of intimacy and privacy. It encourages people to develop, at a very young age, **the ability to understand the digital ecosystem, especially the AIS ecosystem, and to acquire the know-how required to navigate through information, protect our tools and personal data, share content, etc.** This understanding helps create **consent** that is truly free and informed, it also helps us to be able to **contest** algorithm decisions and, eventually, **verify** the relevance of the parameters and data taken into consideration for this decision, when it is justified in intelligible manner. In this sense, digital literacy equips us to understand the digital world and algorithmic decisions, and also provides the ability to act in this world, when faced with these decisions.

[89] UNESCO. (2009). *Belém Framework for Action. Living and learning for a viable future : the power of adult learning.*

[90] From UNESCO. (2015). World Forum on Education, May 19-22, 2015, Incheon, Republic of Korea, quoted in Baril. (March 24 2017). *L'apprentissage tout au long de la vie : définition, évolution, effets sur la société québécoise*. 9e Journée professionnelle de Bibliothèque et Archives nationales du Québec, Montréal. Viewed online at http://www.banq.qc.ca/documents/services/espace_professionnel/milieux_doc/services/journees_professionnelles/apprentissage/Baril.pdf

[91] p. 1. Québec, C. (2018). *Éthique et cybercitoyenneté : Un regard posé sur les jeunes.* Québec: Commission de l'éthique en science et en technologie (CEST). Viewed online at http://www.ethique.gouv.qc.ca/fr/assets/documents/CEST-Jeunesse/CEST-J-2017/CEST_avis_Cybercitoyennete_FR_vf_Web.pdf

In order for this to happen, developing **critical judgment** is necessary, not only to know how to use digital tools and AIS in responsible manner, but also to **know when to trust or doubt** certain sources, recommendations and enticements—even to defy certain types of manipulation or domination. By integrating training on this critical judgment, digital literacy should allow individuals to exercise more freedom in their AIS use, by avoiding having a particular lifestyle imposed on them (see the autonomy principle).

SHOWING RESPECT AND TAKING RESPONSIBILITY

By combining understanding and critical judgment, digital literacy should lead everyone to be accountable for protecting their own privacy as well as that of others (the privacy principle)—without, however, other actors seeing their responsibility reduced in regards to respect for privacy and the autonomy of digital tool and AIS users. This may be a matter or protecting one's personal data, deciding to share it or asking to verify it. It may also mean knowing how to act respectfully towards or through AIS, by not harassing or cyberbullying through digital media. The digital space is a collective living space, and digital literacy must help improve how we live together in this space, while encouraging governments, companies, schools and parents to assume their share of "responsibility in terms of education, awareness and empowerment [...] for the sake of consistency and according to our society's values" [translation][92].

This combination of understanding, critical judgment and respect helps equip people to have their freedoms as users and citizens respected, allows them to participate benevolently in a society that has more and more artificial agents and is linked by digital media, but also to have their voices heard regarding AIS development.

CONTRIBUTING TO THE SUSTAINABLE WELL-BEING OF SOCIETY

Digital literacy can, moreover, help with the response to mental health issues—such as anxiety disorders, mood disorders and dependency problems[93], as well as sustainable development associated with AIS development (the well-being principle).

Regarding mental health, the development of digital literacy should begin as early as possible by limiting the use of digital material in order to reduce the risk of dependency. The basics of algorithmic culture should therefore be taught, as much as possible, using non-digital tools and techniques[94]. Digital education would do well to teach ways of preserving moments of disconnection, to encourage imagination and to manage, or even reduce, stress and anxiety factors generated by digital interactions.

Learning environmentally responsible practices also deserves to be an integral part of digital literacy teachings. This could consist, for example, of making people aware of the high energy costs associated with AIS. This could also mean acquiring creative skills and DIY reflexes to fix objects rather than throw them out, thereby limiting digital waste.

[92] CEST, op. cit., p. 33, *Responsabilité individuelle et sociale des acteurs du numérique*

[93] https://www.jeunes.gouv.qc.ca/politique/habitudes-vie/sante-mentale.asp

[94] CNIL, op. cit., p. 54

3.2.2 Appropriating digital culture: accessibility, inclusion and diversity

DIGITAL INCLUSION

The development of digital literacy raises the issue of a digital divide, consisting of the existence of "inequality in opportunities to access and contribute to information, knowledge and networks, and benefit from the major development opportunities offered by information and communication technologies" [translation][95]. The scale of this divide may depend on accessibility to digital infrastructure (equipment) and the ability to develop the skills and knowledge required to fully use these technologies. Digital literacy should be developed so as to **make the digital world a tool for inclusion,** to be used by anyone, regardless of sex, age, handicap or geographic location.

Given that Canada is unevenly equipped, in terms of infrastructure, to offer all Canadians high-speed Internet access, and that schools, libraries and other community spaces are also unevenly equipped with technology, digital literacy in Canada suffers from an **uneven distribution across the country.** This situation creates a demand for public policies and programs that will **bridge the "digital divide" (geographic and generational)** and the gap between those who have digital skills and those whose level of digital literacy is low.

With this in mind, an intersectorial and interregional round table on digital literacy in Quebec was launched by Printemps numérique in September 2018 to identify "collective action priorities to improve the quality and conditions of digital literacy" [translation][96]. This round table is part of the Jeunesse QC 2030 project, supported by the Secrétariat à la jeunesse du Québec, with a mandate examine the realities of Québec youth regarding the digital world by meeting them at digital cafés held in various cities across Québec[97].

Digital inclusion can also be fostered through digital education given in such a way as to help develop solidarity between people, communities and generations (see the solidarity principle). Intergenerational and peer learning would therefore be worth promoting.

AN ISSUE OF CITIZEN PARTICIPATION

Since it is inseparable from cyber citizenship training, digital literacy becomes a shared responsibility allowing everyone, across the country, to participate in community life, in which digital technologies play an integral part. If citizen participation were solicited commencing at the design phase of certain AIS in order to discuss the social parameters of AIS, their objectives and the limits of their decisions (see the publicity principle), any individual could therefore be included in this discussion and thus take part in the search for creative solutions that are ethically acceptable and socially responsible (see the autonomy principle).

Digital literacy would at the same time be inseparable from digital culture by taking the form of **popular education** through **mediation** initiatives with all population categories across the country[98]. This was suggested not only by citizens involved in the Montréal Declaration (see Part 4, Report on the Results of the Winter Co-construction Workshops, section 5.2), but also in the reports of the CNIL and the IEEE which highlight the importance of raising public awareness around ethical and security issues related to artificial intelligence technologies, both to ensure informed and safe use, but also to reduce fear, confusion and ignorance about the issues raised by these technologies.

[95] Michel Élie. 2001. *Le fossé numérique, l'internet facteur de nouvelles inégalités ?*. Problèmes politiques et sociaux (861) : 33-38. Cited in: Québec: Commission de l'éthique en science et en technologie (CEST). 2018. *Éthique et cyber-citoyenneté: Un regard posé sur les jeunes*. Online: http://www.ethique.gouv.qc.ca/fr/assets/documents/CEST-Jeunesse/CEST-J-2017/CEST_avis_Cybercitoyennete_FR_vf_Web.pdf (p. 14)

[96] https://mailchi.mp/358e547609f8/le-pn-lance-la-premire-table-de-concertation-en-littratie-numrique-au-qubec?e=d4a8cb83f8

[97] http://www.printempsnumerique.ca/projets/projet/jeunesse-qc-2030/

[98] CNIL, op. cit., p. 54.

INCLUSION SPACES: LIBRARIES AND THIRD-PARTY SPACES

Libraries play a key role in digital inclusion and literacy, whether through access to technologies and to quality online information regarding health care, education and work, or by strengthening critical digital skills in a lifelong learning perspective. We can then talk about digital empowerment, or developing abilities that allow us to live, learn and work in a digital society.

Digital inclusion is tied to digital literacy, as it focuses on the politics, services and spaces that aim to reduce barriers to access, facilitate knowledge sharing (in particular local or critical), and ensure the active participation of excluded audiences by making them a priority. In this sense, digital empowerment is a condition of digital inclusion in the context of emerging AIS.

Libraries which integrate empowering and inclusive approaches in terms of access, training, safe spaces—both for physical integrity and exercise of freedom—are designated third-party spaces.

Third-party spaces, whether libraries, fab labs[99], or community or cultural centres, foster trust and engagement through common spaces which are open, flexible and facilitate collective use, and even collaborative design, digital community learning, and democracy-transforming conversations. The "make together" through the creation of social and shared ties amplifies digital inclusion and literacy by contributing to an active citizenship, which ultimately creates "live together".

[99] Or "fabrication laboratories". These are spaces dedicated to building projects through a series of free and open-source software and solutions. http://fabfoundation.org/index.php/what-is-a-fab-lab/index.html

4. DIGITAL INCLUSION OF DIVERSITY PROJECT

Although disagreements around the meaning of democracy are still raw, there is nevertheless a consensus over a democratic ideal: the inclusion of all in a society of equals. Conversely, the exclusion of one part of the population of the political community for economic, social, political, cultural, religious or ethnic reasons, among others, appears as a failure of democracy if the exclusion is not intentional, and as a political mistake if it results in intentional discrimination. The ideal of democracy, whatever its faults may be, and perhaps even because of its failure to overcome them, is contained in the expression "no one should be left behind".

As could be expected, the citizens who took part in the Declaration's deliberative workshops strongly voiced this inclusion ideal and worried that AI may be developed at the expense of part of the population, increase inequalities or cause new discrimination, either directly or indirectly and in an insidious fashion[100]. The problem of discrimination and the inclusion issue were discussed from not only a legal and democracy perspective, but also in terms of knowledge and privacy. Although the principle of justice itself justifies the importance of including diversity and making it one of the purposes of democracy, there exists another instrumental reason: diversity can be sought as a way to improve collective thinking in order to stimulate creativity and innovation. The homogenization of society and its components (economic elites, political classes, researchers, office employees, etc.) usually if not always leads to a loss of creativity and of the ability to adapt to technological and social changes.

The deliberations helped refine our understanding of the issues around democratic inclusion in AI development and helped enrich the Declaration's principles, highlighting the relevance of formulating a diversity inclusion principle that is not simply democratic participation or equity, but one that is closely tied to these issues.

[100] See Part 3 Results report: winter co-construction workshops, Section 4.4

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.

This diversity inclusion principle applied to artificial intelligence systems (AIS) recalls the right to equality and non-discrimination declared by the Universal Declaration of Human Rights (art. 7)[101] and by the various charters of rights and constitutions of democratic societies. Article 10 of Québec's Charter of Human Rights and Freedoms discusses the link between equality, freedom and the right not to be discriminated against; it is worth quoting in its entirety:

"Every person has a right to full and equal recognition and exercise of his human rights and freedoms, without distinction, exclusion or preference based on race, colour, sex, gender identity or expression, pregnancy, sexual orientation, civil status, age except as provided by law, religion, political convictions, language, ethnic or national origin, social condition, a handicap or the use of any means to palliate a handicap.

Discrimination exists where such a distinction, exclusion or preference has the effect of nullifying or impairing such right."[102]

Lastly, under article 15 of the Canadian Charter of Rights and Freedoms:

"Every individual is equal before and under the law and has the right to the equal protection and equal benefit of the law without discrimination and, in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental or physical disability."[103]

Although these ethical and legal principles were shared by the participants in the deliberations of the Declaration's co-construction process, whether they were citizens, experts or stakeholders, and by the different actors in AI development, moving on to recommendations and actions with respect to these ethical and legal standards is not easy and comes up against a series of difficulties. The first one lies in identifying incidents of discrimination and exclusion that could be tied to AIS use. A second difficulty consists in identifying the potential causes of discrimination, and determining the consequences of discrimination on people's autonomy, on their ability to lead a dignified life aligned with their conception of what is good. Another difficulty concerns the understanding of diversity, and can be summed up as follows: Diversity of what? Inclusion in what? We will not provide an *a priori*, overly restrictive definition of diversity. The co-construction process generated discussion of different aspects of diversity that are often studied separately: the diversity of the results produced by AIS, the diversity in AIS's data inputs, the diversity of their users, the diversity in sexuality (gender and sexuality) and of cultural minorities in the development of AIS, etc.

[101] *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence,* CNIL

[102] Charter of Human Rights and Freedoms, 1975, art. 10.

[103] Canada Act 1982, 1982, ch. 11 (UK), art. 15.

Among the results from the co-construction process worth mentioning is the idea that AIS shape the context in which our identity is formed, by reducing the diversity of available options and proceeding by stereotype, thereby deeply affecting our very identities. The second result is that the issue of diversity must not only be understood from the point of view of AIS operations, but rather from the point of view of the social mechanisms that make its development and rollout possible. This is a "social critique" perspective. Stated simply, the research settings for computing and AIS industrial design, among other things, are spaces that are not immune to sexual, social, cultural and ethnic discrimination, and can even help make them worse. These types of discrimination, as we will note below, are rarely intentional, but rather indirect, systemic and not sought out. They are nonetheless significant problems, and reflect deeper, more hidden mechanisms of exclusion or marginalization.

One issue that the co-construction process barely scratched, but that needs to be acknowledged, is the inclusion of diversity in the rollout of AI at the international level. We cannot ignore the fact that AI development is an important economic and strategic issue, subject to intense international competition for which certain nations are structurally disadvantaged and are perceived as predatory spaces (based on cheap IT labour, unprotected data, failing public health care, legal and police services, and natural resources that are already controlled by foreign companies).

4.1

ALGORITHMIC NEUTRALITY QUESTIONED

Human biases and impartial machines?

As soon as you discuss AIS operations and their social interest, you run into a paradox: what is attractive about algorithms (learning or not) is that they allow us to automatically obtain the desired result while eliminating human reasoning errors. Yet the idea that algorithms can also amplify human biases is not unfounded, and tempers the trust we have in algorithmic impartiality. To truly understand this paradox, we must first go back to the assumption that algorithms, and especially those found in AIS, are less biased than humans.

The first thing to consider is that human beings, although gifted with an intelligence more complex than that of algorithms, are quick to make mistakes due to their emotional state[104], level of fatigue and concerns, but above all their cognitive and ideological biases, which are difficult to eliminate. Cognitive biases are intuitive ways of thinking that distort (bias) logical reasoning and lead to erroneous beliefs[105]. Among the approximately forty recorded biases, one should mention confirmation bias, which is the tendency to only seek out information that confirms our beliefs and refuse information that contradicts them. One bias that plays an important role in forming ideological biases and the genesis of direct social exclusions is the negativity bias, under which we remember negative experiences more than positive ones (this bias also allows us to learn from tragic mistakes). Human beings have a tendency to ignore their own biases and not to see them at work in their quick reasoning. This is especially problematic when an urgent decision needs to be made, one that has important repercussions for oneself and others.

The use of algorithms to solve problems or make the best decision in an emergency, with incomplete information and under uncertainty has proven to be of great value. In its most fundamental meaning,

[104] On the different dimensions of emotions in the knowledge and reasoning processes, see Joseph Ledoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life,* New York, Simon & Schuster, 1998. Also see Antonio Damasio's work *The Feeling of What Happens: Body and Emotion in the Making of Consciousness,* New York, Harcourt Brace & Company, 1999.

[105] On cognitive biases, see Daniel Kahneman, *Thinking, Fast and Slow,* Farrar, Straus & Giroux, 2011.

an algorithm is a set of instructions, a recipe built from programmable steps, developed in order to organize and act upon a body of data, in order to quickly arrive at the desired result[106]. The interest of their design and use is twofold: an algorithm helps automate a task and always obtain the desired result; it helps eliminate the biases that affect human reasoning. One of the famous cases that helped reduce the rate of infant mortality at birth is Dr. Apgar's test, which consists of a formula with 5 variables (heartbeat, breathing, reflexes, muscle tone and colour) to evaluate a newborn's health status[107]. With a very basic procedure, Dr. Apgar's formula helped arrive at a better result than human intuition in difficult circumstances for exercising judgment. This is the triage principle used in hospital emergency rooms.

Kahneman (2011) easily convinces us that algorithms are generally more reliable than humans because they are not biased. Of course, it is human beings who design the algorithm based on the result they seek. But the algorithm user only needs to apply it to obtain the correct result. In the case of AIS, the machine engages a learning algorithm capable of identifying patterns in gigantic sets of data, of learning by itself by interacting with its environment, and of applying different lines of instructions. Free of the biases that corrupt human reasoning, AIS are supposed to be neutral tools that provide neutral results.

On this subject, the citizens had seemingly contradictory beliefs. On the one hand, they expect AIS to be more neutral or impartial than human beings, and stated their hope that digital judges will make better decisions. On the other hand, they do not trust them, questioning their impartiality. They were concerned about the fields of justice and predictive policing, but also the health care and human resources sectors. Under the veneer of neutrality, automatic decision-making may hide biases and exacerbate, even create discrimination[108].

Discriminating Machines

Although one can nurture fears around AIS, it is not easy to demonstrate whether they are biased and say which ones are, or what the causes are. In the Declaration's consultation process, the participants were presented with a scenario designed to spark discussion. The algorithmic biases and resulting discrimination were clearly identifiable. Outside of this context, it is not easy to identify the discrimination or marginalization effects caused by algorithms, and even harder to correlate them with algorithmic biases. However, a critical analysis of AIS operations and a tracing of the socioeconomic paths of vulnerable individuals and populations helps establish some correlations between AIS use and certain types of discrimination.

Recent work by Virginia Eubanks[109] has helped document specific cases of algorithmic discrimination. In a book with a very evocative title, *Automating Inequality*, Eubanks rigorously studied the automated systems that determine which people are eligible for social benefits and medical reimbursements and which ones are no longer eligible. Eligibility can be determined by a set of criterias that includes current financial situation, data on housing and area of residence, health status, etc. With the arrival of computers, databases have grown and both public administrations and private companies (banks, insurance companies) have access to them and can process historical data: Does the person have a medical history? Since when? How many times have they needed medical care? Have they always repaid their credit on time? With the development of AIS, not only are we processing much more data to refine the profiles of clients, but we can also make predictions about their behaviour, their solvency or changes in their health. Indeed, one of the virtues of AIS, which explains in part their massive rollout by administrations and private companies, is this ability to make increasingly rich and often very precise predictions. One of the reasons for their success is that human beings

[106] Tarleton Gillespie, *Algorithm," in Digital Keywords: A Vocabulary of Information Society and Culture*, dir. Ben Peters, Princeton, Princeton University Press, 2016. Preliminary version available online: http://culturedigitally.org/wp-content/uploads/2016/07/Gillespie-2016-Algorithm-Digital-Keywords-Peters-ed.pdf

[107] Kahneman (2011), chap. 21 *Intuitions vs. Formulas*; Atul Gawande, *A Cheklist Manifesto*, New York, Metropolitan Books, 2010.

[108] See *Bots at the Gate* report, The Citizen Lab, University of Toronto, p. 31. https://ihrp.law.utoronto.ca/sites/default/files/media/IHRP-Automated-Systems-Report-Web.pdf (p.31)

[109] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press

are predictable enough in their behaviour, and the reasons behind their habits are easily detectable by a well-designed AIS.

But what this prediction function also makes possible is a profiling of people to avoid taking any risks that could result in a cost to the administration or private company. As soon as an algorithm identifies a risk related to a person's profile, it also launches closer surveillance processes or exclusion from social assistance programs, health insurance, recruitment, etc.

Simple scoring systems, which were the very basis of Dr. Apgar's formula that helped save lives, also tend to automate exclusion and inequalities by systematically flagging poor or vulnerable people as being at risk. As Virginia Eubanks demonstrates, these automated systems have a tendency to punish poor and marginalized people. In fact, by flagging them as being at risk, AIS expose them to added risks of marginalization[110]. Through a feedback loop, these prediction tools are likely to create the difficulties they claim to be flagging[111]. For example, an automatic recruiting system based on scoring applicants at a hiring interview will learn to reject those who present a risk of absenteeism, or of poorer workplace performance, because they live far away from their future workplace. Yet this type of decision, which discriminates against candidates according to their place of residence, can reinforce socioeconomic inequalities. This is exactly what happened in the case of the Xerox company, as documented by Cathy O'Neil[112]. The people whose applications were rejected lived in far away residential areas... and were poor. With lower scores because of a financially disadvantaged environment, these people had fewer chances of finding work and were more at risk of job insecurity. In the case of Xerox, the company noticed this discriminatory result and modified the algorithm's model: "The company sacrificed a bit of efficiency for fairness."[113]

More and more problem cases are being reported: predictive calculations seem to reproduce or accentuate exisiting inequalities and discrimination in society. Amazon's algorithm, for example, was treating clients differently according to their place of residence, and for unknown reasons (as the algorithm cannot be accessed), did not offer same-day delivery to people in predominantly African-American neighbourhoods[114]. In the field of justice, algorithms are increasingly used to predict the risk of recidivism. The interest in crime prediction comes from the fact that both the prison population and the cost of imprisonment have greatly increased; a better prediction of risk of recidivism allows inmates with a low risk of recidivism to be set free or, in other words, it frees up room in prison. In 2016, the ProPublica website's investigation showed that the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithms from Northpointe, Inc., used by the Florida justice system, predicts the risk of recidivism among black criminals as twice as high as the risk among white criminals[115].

Surprisingly, we could say more succinctly that AIS are victim to biases similar to cognitive biases, such as confirmation bias: the discriminatory treatment of certain groups not only reinforces inequality, but maintains the conditions for social violence. By predicting that African-American criminals are twice as likely to reoffend, thereby increasing the rate and length of incarceration for this population, AIS tend to create a serious discrimination situation, or at least perpetuate it. And the discrimination machine is self-perpetuating, only looking through the data to find what confirms its own predictions.

We could object that AIS are not the source of the problem, that discrimination has always existed and that algorithms are "neutral" tools for policies that are anything but. This objection is not unfounded. It reminds us that we must distinguish the tool (AIS) from its use (a discriminatory policy).

[110] Citron, D., and Pasquale, F. *The Scored Society: Due Process for Automated Predictions.* 89 Washington L. Rev. 1, 2014. https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf?sequence=1

[111] Michael Aleo & Pablo Svirsky, Foreclosure Fallout: *The Banking Industry's Attack on Disparate Impact Race Discrimination Claims Under the Fair Housing Act and the Equal Credit Opportunity Act*, 18 B.U. PUB. INT. L.J. 1, 5 (2008).

[112] Cathy O'Neil (2016), chap. 6 *Ineligible to Serve: Getting a Job*.

[113] Cathy O'Neil (2016), p. 119. *La compagnie a sacrifié un peu d'efficacité pour plus d'équité*.

[114] *Amazon same-day delivery less likely in black areas, report says*, USA Today, April 22, 2016: https://www.usatoday.com/story/tech/news/2016/04/22/amazon-same-day-delivery-less-likely-black-areas-report-says/83345684/

[115] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, 23 May 2016, *Machine Biais*: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

A critical examination is required, however, of the tool itself and its practical applications. First, when they are developed for certain policies such as evaluating recidivism, the tools produce some of the discrimination mentioned above and can no longer be considered "neutral". Then, algorithms are not infallible and their reliability is very relative, depending on the field and the mathematical model used[116]. As the Propublica journalists observed in the May 23, 2016 investigation, although the COMPAS algorithm gives more reliable results than chance for all crimes taken together, it gives incorrect results for violent crimes (those that do lead to longer sentences). We could be satisfied with the fact that, overall, the COMPAS algorithm is more reliable than chance, but in a democracy that recognizes each person's right to be treated fairly, this is not relevant: if overall the algorithm is reliable, it sacrifices the fundamental interests of too many people for its use to be legitimate.

Lastly, let us add that implementing AIS reduces the opportunities for appeal, as AIS are considered, wrongly, to be very reliable and unbiased. Virginia Eubanks's personal story is instructive: when confronted with a decision made, in all likelihood by an algorithm, to suspend her medical coverage, she was able to rely on her knowledge of algorithm operations, her employer and her material resources.

The cases we have just discussed all occurred in the US. But Canada should beware of the predictable consequences of AIS use by Canadian public administrations and learn from the unfortunate experiences in other countries. Although automation has considerable appeal for the processing of millions of files that traditional administrations can hardly handle, the risks of violating the fundamental rights of citizens are sometimes too great. The case of processing immigration files is a strategic issue for Canada. Hundreds of thousands of people come into Canada each year for very different reasons and seek to obtain temporary or permanent resident status. Studies led by the University of Toronto's Citizen Lab highlight the impacts of automated decision-making on immigration requests and the way the technology's mistakes and assumptions may lead to serious consequences for immigrants and refugees[117]. The complexity of many immigration requests, in the case of political refugees, for example, could be inappropriately handled by AIS, leading to serious violations of human rights protected by various international conventions that Canada has signed. The ethical principles of the Declaration and Quebec, Canadian and international law suggest that precautionary measures should be taken with AIS, which have the potential to cause serious discrimination.

Biased Identity: the Internet and AIS

The AIS used by the vast majority of the population are inseparable from the most basic Internet operations: they are the classification and recommendation algorithms (used by Google, Amazon, Spotify and Netflix) as well as the social networks (Facebook and Twitter, for example). In every case, algorithms learn from the tracks that Internet users leave behind signalling their regular behaviour, their preferences and tastes, their political ideas and their worldviews. On the one hand, their searches on the web and their social media interventions, whether verbal or non-verbal (posting pictures online), say something about their "me", their identity, and on the other hand, Internet users build representations of their identity based on their intended audiences[118]. These representations are consumer goods for social media audiences, but more widely and more authentically for the algorithms of online companies that gather data to sell products, goods and services, either to individuals or other companies: the data itself or the space for targeted advertising[119]. Yet algorithms represent other intermediaries, free agents that shape the representations and identities of users.

[116] Crawford, K. and R. Calo, *There is a blind spot in AI research,* Nature, 20 October 2016, doi: 10.1038/538311a

[117] https://ihrp.law.utoronto.ca/sites/default/files/media/IHRP-Automated-Systems-Report-Web.pdf

[118] Lee Humphreys, *The Qualified Self: Social Media and the Accounting of Everyday Life,* Cambridge, The MIT Press, 2018.

[119] Cathy O'Neil (2016), chap. 4 *Propaganda Machine: Online Advertising*.

In line with the academic studies on the workings of ranking algorithms and social media, the participants in the Declaration's co-construction process raised the issue of the influence of AIS on cultural diversity and the identities that tend to both be segmented into groups and homogenized within each group. To better understand this phenomenon, we must change our view of algorithms and define them, as Lessig (2006)[120], Napoli (2014)[121] or Ananny (2016)[122] do, as governing institutions: "Code is Law," said Lawrence Lessig, Harvard law professor and pioneer of the commons movement. In other words, software programs constitute law. Indeed, algorithms have the power to structure behaviours, influence preferences, guide consumption and produce consumable content for prepared, even conditioned Internet users. This power is therefore being exercised on the very identity of Internet and connected object users, and biases this identity by shaping it.

By ranking the contents and making recommendations, algorithms more fundamentally have an ability to "structure the possibilities" offered to users[123] and create a digital universe where search and information pathways are mapped out. The ranking and filtering of information that has become overabundant will indirectly harm pluralism and cultural diversity: by filtering the information, by relying on the characteristics of their profiles, algorithms will increase the tendency among users to frequent people and seek content (in particular, opinions and cultural works) that are *a priori* aligned with their own tastes, and reject the unknown[124]. An individual is then trapped in a "filtering bubble", that is to say a set of recommendations that are always in line with the profile he or she is developing through digital behaviour and which is encouraged by the digital environment that is adapting to it. The effects of an unprecedented boom in content and cultural offerings are paradoxically neutralized by a phenomenon of effectively reduced individual exposure to cultural diversity. And this occurs even if the individual wants such diversity.

An objection could be raised here: what algorithms make possible is the personalization of user profiles that, because of the diversity of people, effectively increase the diversity of offerings. This objection could be serious if algorithms did not favour popular content and did not guide searches and recommendations to showcase this content. This is reinforced on social media through the well-known phenomenon of polarization, which affects how opinions and groups are formed[125]. The way social networks operate accelerates polarization in two ways:

1. first because apps provide users with tools that allow them to filter the news according to their interests and the people they connect with, based on personal affinities. The famous Twitter #hashtag is probably the most effective filtering tool; Cass Sunstein discusses the "hashtag nation" in #republic (2017)[126], and
2. second, the algorithms of these social networks learn to spot what matters to users and only gives them information that they are supposed to be interested in. By cross-referencing this with personal data left behind on other websites, algorithms build a powerful echo chamber in which the same people, according to their apparent interests, are put in touch with each other, "connect", exchange converging viewpoints, reinforce their beliefs and consolidate their collective characteristics.

Consequently, even if a wide diversity of groups, newsfeeds and profile recommendations are generated by social media algorithms, this diversity is a facade: not only does the internal composition of such groups tend to homogenize, but the groups

[120] Lawrence Lessig, Code: *And Other Laws of Cyberspace, Version 2.0*, New York, Basic Books, 2006.

[121] Philip M. Napoli, *Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption*, Communication Theory 24 No. 3 (2014): 340-360. In particular, the *Institutionality and algorithms* section, p. 343 and following pages.

[122] Mike Ananny, *Toward an ethics of algorithms: Convening, observation, probability, and timeliness,* Science, Technology, & Human Values 41, No. 1 (2016): 93-117..

[123] Ananny (2016): *Algorithms 'govern' because they have the power to structure possibilities*, p. 97.

[124] See CNIL report, *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence,* 2016.

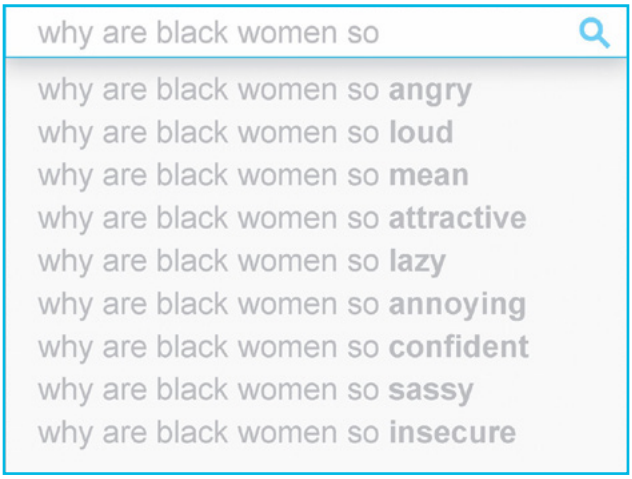
[125] See the many works of Cass Sunstein on the subject, for example: *Infotopia*, Oxford, Oxford University Press, 2006.

[126] Cass Sunstein, *#republic*, Princeton, Princeton University Press, 2017, p. 79.

remain relatively impervious to one another. AIS operations therefore separate individuals who are different and bring together individuals who are similar. The inclusion of diversity calls instead for an inclusive diversity: different people gathered to exchange and learn from each other's differences.

To achieve this goal, representations of socially disadvantaged groups or practising minorities (cultural, religious, sexual) should, at the very least, not be caricatures or stigmatizing. That requirement has not been met. Academic studies are unanimous: ranking and recommendation algorithms are not neutral and reflect the biases currently found in society. More specifically, they recreate the social structures of domination and exclusion and help reinforce them. This is what Safiya Umoja Noble very clearly demonstrates in her reference book, Algorithms of Oppression (2018)[127] by specifically examining how the Google Autocomplete algorithm operates[128]. The book's cover illustrates the problem (see Figure 1).

Figure 1: Detail from the cover of Safiya Umoja Noble's book, Algorithms of Oppression



The search "Why are black women so..." generates the following suggestions: "... angry", "loud", "mean", "attractive", "lazy", etc. Without going into a detailed analysis, it is clear that Google's Autocomplete algorithm suggests negative representations of black women that stigmatize them. Open searches such as: "black women" generate suggestions for pornographic websites, reducing black women to sexual objects[129]. This reinforces cultural stereotypes[130] and discourages people from making unpopular searches[131].

This type of recommendation is problematic for at least two reasons: it projects a tarnished image of a stigmatized group to society and helps maintain the symbolic conditions of domination on this group, by reinforcing stereotypes. Furthermore, it reflects a tarnished image to the members of the represented group and affects their foundation of self-respect, their sense of self-esteem and their confidence in their worth. This submission or subjection to representations of self that are defined by others is a major factor in domination by others. The examples of identities biased by algorithms are too many to list. To conclude with a more subtle example, consider the case of a Google translation from Turkish to English:

O bir doctor / O bir hemsire.

The same neutral turn of phrase in Turkish, with an undetermined personal pronoun, is translated two different ways in English, associating the role of a doctor with being a man and the role of a nurse with being a woman: "He is a doctor," "She is a nurse."[132] In this case, the problem is the gendered allocation of social roles and professions, which, incidentally, regardless of their respective importance and merit, are a throwback to a hierarchal domination structure in which man commands and woman obeys.

[127] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, NYU Press, 2018.

[128] Garber, M. 2013. *How Google's Autocomplete was... Created / Invented / Born.* The Atlantic. Accessed March 3, 2014.

[129] Safiya Umoja Noble (2018), p. 19.

[130] Baker, P., and A. Potts. 2013. *Why Do White People Have Thin Lips? Google and the Perpetuation of Stereotypes via Auto-complete Search Forms*. Critical Discourse Studies 10 (2): 187-204. doi:10.1080/17405904.2012.744320.

[131] Gannes, L. 2013. *Nearly a Decade Later, the Autocomplete Origin Story: Kevin Gibbs and Google Suggest*. All Things D. Accessed January 29, 2014.

[132] Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCIENCE 183, 183-84 (2017); Calo, Ryan. 2017. *Artificial Intelligence Policy: A Primer and Roadmap.* Washington University. SSRN: https://ssrn.com/abstract=3015350

4.2

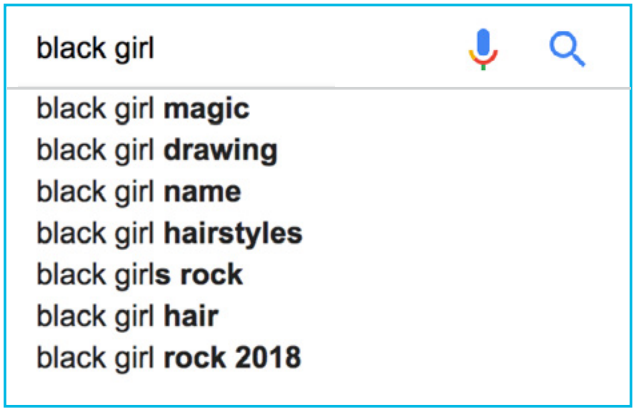
UNBIASING ARTIFICIAL INTELLIGENCE SYSTEMS

If current AIS operations are not neutral and help reproduce the social structures of marginalization, stigmatization and domination, we have to ask how can we fix the situation and reduce the inequalities it causes? We have to state from the outset that the neutrality of algorithms is not the problem that needs to be solved, regardless of what the literature on this subject would have you believe. The ideal is not algorithm neutrality, or at least, algorithms operating neutrally is not enough to satisfy the diversity inclusion requirement in society.

Regardless of the meaning we give to neutrality, it does not allow us to correct what appears to be unintentional discrimination, unless intentions are ascribed to AIS or we demonstrate bad intentions on the part of the incriminated algorithm's designers and developers. If a tool is considered neutral when its use does not affect the state of society, and leaves it intact, then we can see that this is not what we are looking for to correct discrimination, because in fact we are trying to change society. If we admit, instead, that neutrality refers to the use of a tool that does not promote a conception of what is right and is not intended to create an unfavourable situation for part of the population, we are still not addressing the problem. Indeed, the AIS have no "intention" of recreating or reinforcing discrimination and were not developed for that purpose, but they do so on a massive scale because of operational biases (the mathematical model or training data).

It is therefore time to abandon this idea of neutrality, which is not relevant at this level of reflexion. And the reason is not that neutrality is unattainable, but that it is not desirable in AIS design. Rather the critical examination of AIS has revealed that their operations must be corrected in order to avoid recreating discrimination and reinforcing conditions for the marginalization or exclusion of people and groups, according to the social justice and equity criteria applied to human actions. These corrections are possible if humans (programmers, data explorers) get involved. This is what Cathy O'Neil has shown with the Xerox example, since the recruitment algorithm was modified to no longer reject applications from people living in underprivileged neighbourhoods. It is therefore worth mentioning that the situation is improving due to the alerts that are raised regularly and interventions by human beings. As a case in point, the "black women" search provided by Safiya Umoja Noble no longer produces the same results (see Figure 2).

Figure 2: Search on google.com engine performed on October 29, 2018



Much work remains to be done, as Figure 3 illustrates below.

Figure 3: Search performed on google.fr engine on October 29, 2018

How can AIS be unbiased and their development made more inclusive? The answer to this question is not only technical, but also ethical, social and political, and demands that we examine how AIS operate.

A Problem With Data

The first source of bias that stands out when investigating discrimination is the development of the databases used by algorithms. Digital data are like a natural resource that must be extracted, filtered and transformed. Nowadays, the term used is "data mining" (data exploration and extraction); data is compared to oil. There is one fundamental difference, however: unless one refuses all realism, one must recognize that natural resources exist even if we cannot extract them, and even if we cannot see them. Digital data, on the other hand, does not exist without a device to capture and process them. A beating heart is not data; a heart rate captured by a smart watch is data. And even then, that data is not raw because the monitoring device (the heart rate monitor) must be coupled to interpretation devices that produce a measure. Data must be generated and interpreted[133].

Algorithms create associations by detecting and combining the aspects of the world (characteristics, categories of data sets) that they have been programmed to see[134]. There are two types of problems with data: their quality and their extension. The quality of data can be adversely affected by inadequate or morally inappropriate labelling. As it is human beings who must label most training data themselves, human biases like cultural assumptions are also passed on through the choice of classifications[135]. Kate Crawford maintains that we must then adopt a rigorous quantitative approach to examine and evaluate data sources. Even if the methodologies of social sciences can make understanding big data even more complex, it could give the data more depth[136].

Tay, the GIGO phenomenon

Tay is a chatbot created by a Microsoft technological development team. On March 23, 2016, this chatbot was launched on Twitter for the purpose of interacting with other users by processing the messages it receives and publishing messages of its own. The experiment was meant to confirm that AIS could now pass the Turing test, and it was a catastrophe. Tay was "unplugged" less than 48 hours after being launched.

Tay's destiny teaches us something about how algorithms work. By educating itself through interactions with other Twitter users, Tay had very quickly published heinous, racist and sexist messages. Had it been a human being publishing that type of message, he or she would quickly have been called racist and sexist. Tay's behaviour can be explained by the fact that the messages it was receiving were overwhelmingly of a racist and sexist nature. By learning from incorrect data (morally incorrect, in this case), the Tay algorithm gave morally incorrect results. This only confirms a popular expression in the computing world: "Garbage in, garbage out" (GIGO).

The extension of data is the other problem that must be confronted. By this, we mean the fact that the data does not always cover the entire phenomenon that we wish to observe, or there is too much data for a small part of the observed phenomenon. Indeed, one of the meanings of bias is statistical and refers to the gap between a sample and a population. Selection bias occurs when certain members of a population have a greater chance of being sampled than others.

[133] Lisa Gitelman (ed.). 2013. *Raw Data is an Oxymoron*. Cambridge: The MIT Press.

[134] Mike Ananny. 2016. *Toward an ethics of algorithms: Convening, observation, probability, and timeliness.* Science, Technology, & Human Values 41(1): 93-117

[135] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker et Kate Crawford. 2017. *AI NOW Report.* AI Now Institute at New York University; Kate Crawford. 2013. *The Hidden Biases of Big Data* Harvard Business Review 1. See the report of the Big Data Working Group, under President Obama's Executive Office. 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*

[136] Kate Crawford. 2013. *The Hidden Biases of Big Data.* Harvard Business Review 1;Adam Hadhazy. 2017. « Biased Bots: Artificial-intelligence Systems Echo Human Prejudices », Princeton University. https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices

Although this can be explained by human biases in preparing and exploring the data, the most relevant reason is often that systematic inequalities in society are such that one population is overrepresented in the training data, and that, conversely, another population can be underrepresented[137]. Therefore, the data on which an algorithm trains can be biased or false, and present a non-representative sample that was poorly defined before use[138]. A good example is AIS facial recognition: the more white faces there are in the training data, the better the system will perform for that part of the population[139]. On the other hand, as soon as the white population is overrepresented, other populations, such as African-Americans, are thereby underrepresented. The result is then very problematic and there is a tendency to confuse faces, and even associate human faces with the faces of monkeys, such as occurred in the very unfortunate incident in which the Google algorithm tagged black people as gorillas[140].

This phenomenon becomes dramatic in the legal system. In the United States, where different types of AIS are already used to predict recidivism, the main problem, aside from the poor quality of the data, lies in a lack of relevant data[141]. Indeed, if the crimes of one segment of the population (let us say African-Americans) are better documented and archived than the crimes of another segment of the population (let us say white people), the first will be more heavily penalized than the second, thus feeding a "cycle of discriminatory treatment"[142]. This was the problem encountered in a predictive policing tool like PredPol, which was designed according to a mathematical model developed for earthquake risk, but which works with a non-representative set of data.

Making Algorithms Talk

Athough discrimination can be explained for the most part by faulty data collecting and extraction of discrimination, it is also due to the algorithm itself, its code and its mathematical model. Algorithms, unlike computers (computing infrastructure), are not universal in the Turing sense, meaning that they only carry out the task for which they were designed and have objectives defined by their programmers; a computer is a universal machine in the sense that it can accomplish various tasks, but also requires different specialized algorithms for this purpose. This is why we believe that the AIS that produce discrimination consequences are also to blame. For a given set of data, two algorithms with different parameters, mathematical models and objectives will generate different sets of results. We saw this in the Xerox example.

Let us imagine that in order to avoid the stigmatization of target populations by ranking and recommendation algorithms, we agree on the following objective: for a given search, the algorithm should not always return the same results (in a period during which it is not updated). For example, when we conduct a search for "black women", we should not be given pornographic recommendations, nor should we always see the same recommendations for "hair" and "long hair", which have replaced the degrading suggestions, but also build stereotypes. We can then imagine the introduction of a "chance" parameter, a random parameter in the algorithm. By proceeding in this manner, we also solve the problem of filtering bubbles, which have an effect on the diversity and identity of users who are locked inside a user profile.

[137] Artificial Intelligence: Human Rights & Foreign Policy Implications

[138] Neural Information Processing Systems (NIPS): Kate Crawford, 2017. Viewed October 1, 2018, < https://www.youtube.com/watch?v=fMym_BKWQzk >.

[139] Calo, Ryan. 2017. *Artificial Intelligence Policy: A Primer and Roadmap*. Washington University. SSRN: https://ssrn.com/abstract=3015350

[140] Barr, A. 2015. *Google mistakenly tags black people as "gorillas," showing limits of algorithms.* The New York Times.

[141] Matt Ford, *The Missing Statistics of Criminal Justice,* The Atlantic, May 31, 2015 http://www.theatlantic.com/politics/archive/2015/05/what-we-dont-know-about-mass-incarceration/394520/

[142] AI for the Common Good, https://weforum.ent.box.com/v/AI4Good?platform=hootsuite

SETTING UP A SERENDIPITY PARAMETER

The word serendipity was coined by the British writer Horace Walpole, in 1754[143]. The term refers to the act of making a useful discovery by accident, without looking for it. Some of the greatest scientific discoveries, like penicillin discovered by Alexander Fleming, were made by accident. But serendipity is not just a matter of chance; it is the possibility of making an accidental discovery and must be facilitated by an institutional structure: for example, giving researchers time, favouring meetings, not exercising too much pressure[144] on publishing, which takes up research time, etc. Similarly, recommendation algorithms are architectures of choice that may or may not leave room for fortuitous paths to discovery.

No one expressed this link between architectures (of choice) and fortuity better than the author Umberto Eco. In his speech on libraries, delivered in Milan in 1981, he said:

"In a library where everyone circles about and helps themselves, there are always books lying around that haven't been replaced on the shelves [...] This is my type of library, I can decide to spend a day there in the purest joy. I read the newspapers, I bring books to the bar, then I go get more, I make discoveries. I had gone in to tend to, let's see, English empiricism, and instead I find myself among Aristotle's commentators, I get off on the wrong floor, I enter a section I hadn't planned on visiting, medicine for example, and all of a sudden I come across works dealing with Galien, with philosophical references. In this sense, the library becomes an adventure."

If the parameter is known and its impact can be measured from tests, then that would be an algorithm that avoids filtering bubbles and discrimination without having to correct, after the fact and for less than obvious reasons, the results of the algorithm. Take for example Safiya Umoja Noble's search: "Why are black women so...". Today, Google no longer suggests the "lazy" response. Yet, it could also be as useful to come across a recommendation to a page where, instead of a list of links to racist publications, we would see a link to Paul Lafargue's *The Right to Be Lazy,* published in 1883. Putting chance back into the equation and fostering serendipity, although it may seem contrary to the goals of algorithmic programming, is perfectly aligned with the objective of fighting stereotypes. We also find this idea explicitly stated by the inventor of Twitter's #hashtag, Chris Messina[145].

[143] For the history of this concept, see Merton, R. K., & Barber, E. (2004). *The travels and adventures of serendipity: A study in sociological semantics and the sociology of science*. Princeton, NJ: Princeton University Press.

[144] Umberto Eco, *De Bibliotheca*, transl. from Italian by Eliane Deschamps-Pria, Caen, l'Echoppe, 1986.

[145] Quoted by Cass Sunstein (2018), p. 79.

To ensure the algorithms aren't biased, they must be neither black boxes nor silent boxes. Saying "black boxes" signals the fact that the code for private algorithms is inaccessible, hidden, kept secret by the companies that develop them. One of the reasons is that the algorithm is a "secret recipe" crucial for their business and that this is an issue of intellectual property[146], which we admit is true[147]. But the idea of a black box has another connotation: it may be that companies simply do not want to be held responsible for algorithms that cause discrimination. For businesses, the most effective way to protect their business model is to say that the details of algorithm operations cannot be understood, and that if an unfortunate result has occurred, it could not have been foreseen or prevented. Presented as black boxes, algorithms are protected from any outside investigations of the company that develops or uses them. It is understandable that this can inspire fears and fantasies regarding manipulation by private companies[148]. While individuals are increasingly transparent with companies and governments, the technology that makes this possible is becoming increasingly opaque.

Yet, if we can accept that companies do not want to publicly disclose the codes, it is more difficult to understand why the algorithms are not accessible to competent authorities, whether public or public-private. When discrimination affects a person's fundamental rights, the public authorities actually have an obligation to investigate and sanction. Moreover, in the case of public algorithms, a consensus is emerging that their code should be open and accessible.

These black boxes are also "silent" in the sense that they offer users and people subjected to algorithmic procedures no information on AIS operations, objectives and parameters, nor any justifications for the decisions made, or strongly influenced, by AIS. This silence from AIS, or the people responsible for their design and development, is especially problematic in a democratic society that promotes inclusion and justification. At least that is how the participants in the Declaration co-construction process felt, and this reflects a concern among most researchers in ethics and the social sciences. One citizen suggested, for example, that we should always be able to request an understandable explanation for a decision. Stakeholders such as the Ordre des ingénieurs du Québec also called for making algorithmic decisions easier to understand.

Making algorithms more transparent implies three things:

1. that algorithm designers understand how they work (this may appear trivial, but this condition helps counter designer disempowerment strategies);

2. that the designers and developers are able to formulate the algorithm's parameters and objectives in a language understandable to educated people, but not specialists, and that they do so; and

3. that the companies that develop or use an algorithm regularly publish reports on their societal impact (in this case, on the way it affects disadvantaged and precarious groups).

Since SAI algorithms are very complex and their behaviour is difficult to understand, even for specialists[149], researchers have agreed to call for the implementation of testing procedures that would help evaluate the results and eliminate undesirable results *ex post*. This also implies that audits can be performed before an algorithm is marketed and commissioned[150].

[146] Cathy O'Neil (2016).

[147] Yet some criticize the intellectual property and professional standards that keep algorithms private, and demand transparent codes. See Mike Ananny (2016).

[148] On this subject, see Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, Harvard University Press, 2015.

[149] Algorithm complexity must also not be exaggerated for its designers, which contributes to the perception that they are impenetrable black boxes, as Taina Bucher (2018) reminds us. Taina Bucher, *If... Then. Algorithmic Power and Politics,* Oxford, Oxford University Press, 2018, p. 57.

[150] See Cathy O'Neil (2016); AI NOW (2017); National Science and Technology Council & Office of Science and Technology Policy (2016) *Preparing for the Future of Artificial Intelligence*.

Representation and Inclusiveness

To ensure inclusive AI, we must not only be interested in the design and training of the algorithms, but also the material conditions under which they are developed. In particular, there is a need to examine the possible social discrimination that affects (or is produced by) the AI research and industrial development community. There are two reasons to be interested: one is instrumental, and the other ethical.

The first reason to justify the objective of including diversity in the AI development community is that diversity is a condition favourable to scientific and technological innovation. A homogeneous environment is a factor for scientific and intellectual conservatism in general. There is no need to develop this argument here; it has been made by an author such as John Stuart Mill, a case for the epistemic and moral virtues of diversity. It is also one of the reasons why an open and deliberate process was chosen to develop the Montréal Declaration for Responsible AI. But before moving on to the ethical reason, it should be added that inclusion of diversity in the AI community also helps raise awareness among AIS developers of inclusion and discrimination issues. Indeed, one of the explanations for AIS biases that we have, for the moment, set aside, is the biases of the programmers themselves. It must be said that the vast majority of AI researchers and developers are men. In a North American context, it must be added that they are white men, well paid, with very similar technical educations[151]. One could surmise that their interests and life experiences influence their design and programming of algorithms[152]. A balanced representation of the diversity in society is not a guarantee that algorithm development will be less biased, but it nonetheless would appear to be a mandatory requirement.

If the instrumental reasons for fostering inclusive AI development are important and should be enough to motivate businesses, research centres and universities, the ethical reason is an imperative of a higher order. It is a question of social equity.

We will only be concerned with the case of the presence of women in the AI environment, for brevity's sake, but the study should include an examination of the situation of ethnic and cultural minorities. We observe that women are statistically less present in new digital technologies in general and in AI in particular. This could be explained by the fact that women are less interested than men in computer science. Obviously this answer would be insufficient, because then an explanation would be required for why they are less interested than men in computer science. The most credible hypothesis is that women are less present than men in the field of computing today not because of a lack of interest, or even a lack of training, but because of strong competition with men to earn a place in a social sector that is highly valued and rewarded. This competition is biased from the outset by the fact that women are discouraged from entering it.

It is hard to corroborate this hypothesis in this programmatic chapter on inclusive AI development. However, many studies show that women are the victims of distorted competition that favours men. We will simply quote two examples to end this chapter. The first comes from the British history of AI, which was remarkably recounted in Marie Hicks's book with the eloquent title: *Programmed Inequality*[153]. Marie Hicks demonstrates that the United Kingdom, in the wake of the Second World War, had a class of workers in the computing sector where the ratio of women was very high. Computing jobs were low paying at the time. But starting in 1964, these jobs became more valued and the British government committed the country to a technological revolution. Marie Hicks notes that at the same time, the image of women was being used to advertise and sell machines, and that computing jobs gradually became considered for men. The role of manager became emblematic in this technological revolution and was associated with men. This is how women were pushed aside from the most valued computing jobs.

The second example completes the first and illustrates the vicious cycle between algorithmic biases and discrimination based on sex in the field

[151] For statistics in a U.S. context, see the U.S. Equal Employment Opportunity Commission's report, *Diversity in High Tech* (2016).

[152] Safiya Umoja Noble (2018)

[153] Marie Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*, The MIT Press, 2017.

of AI development. A study by Carnegie Mellon University, conducted by Amit Datta, showed that on Google, women had fewer chances than men of being targeted by ads for high-paying jobs (US$200,000)[154]. As Kate Crawford remarks, if women do not have access to these ads, how can they apply for the jobs[155]? Knowing that AI jobs are now very well paid, the risk is high that women will be discriminated against from the moment the position is posted. This situation needs to be urgently addressed to ensure that the social development of AI is truly inclusive.

[154] Amit Datta, Michael Carl Tschantz, and Anupam Datta, *Automated Experiments on Ad Privacy Settings.* Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

[155] Kate Crawford, *Artificial Intelligence's White Guy Problem*, New York Times, 25 June, 2016. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0

5. ENVIRONMENT PROJECT: AI and environmental transition, issues and challenges for strong sustainability

Many of the citizens who took part in the Montréal Declaration deliberative workshops felt strongly that AI must be developed in a way that is sustainable for the planet. Indeed, given the current state of the environment, with the global climate change crisis, the energy transition, the accelerated depletion of natural resources and the collapse of biodiversity, many environmental issues were raised around the digitizing of society, including data storage. Some citizens spoke of outrageous accumulations of data and the related energy costs, or the massive and catastrophic accumulation of data in the worldwide cloud. There was also the issue of electric and electronic waste, and the planned obsolescence of electronic objects in our everyday lives.

Other participants also highlighted the potential contributions of AI to environmental management, for example by automatically monitoring lands that are rich in biodiversity. They also discussed the fact that applications made possible by AI, such as self-driving cars, should not be used at the expense of active mobility (walking, cycling), which holds more promise for the ecological transition of cities. Lastly, during the last deliberative workshop in October 2018, a team worked directly on a prospective scenario of algorithmic governance of individual behaviours and the environmental rebound effects. This discussion group listed many ethical and democratic issues that must be resolved to guide such an initiative.

These discussions thereby helped highlight the importance of the environmental issue in the global development of AI, and helped enrich the Montréal Declaration's principles. The relevance of formulating a new environment principle appeared inescapable.

SUSTAINABLE DEVELOPMENT PRINCIPLE

AIS must be developed and used so as to ensure strong environmental sustainability for the planet.

This requirement for strong sustainability underscores the fact that AIS deployment and its effects on society must be compatible with the planet's environmental limits, the pace of resource and ecosystem renewal, climate stability and the non-substitutability of natural assets by artificial assets[156].

The European Group on Ethics in Science and New Technologies, in its paper Statement on *Artificial Intelligence*, Robotics and *"Autonomous" Systems* (2018)[157], defines nine ethical principles and democratic prerequisites, with the ninth one addressing sustainability. This principle also tends towards a logic of strong sustainability by recommending support for "the basic preconditions for life on our planet", the "preservation of a good environment for future generations", as well as "the priority of environmental protection".

This document expands upon these environmental issues of AIS. First, it addresses the issue of the current contradiction between the digital transition and the environmental transition. Then, it clarifies this issue from an artificial intelligence standpoint by distinguishing what relates to the AI's environmental footprint, with the environmental effects it brings, from AI as a tool in the service of the environmental transition. This report on priority actions concludes with recommendations for strong sustainability for AI systems in society.

[156] For an overview of this concept, see: Bourg D. and Fragnière A. (2014), *La pensée écologique. Une anthologie*, Article : *Enjeux économiques : durabilité faible ou durabilité forte*, p. 439-443.

[157] https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

5.1

DIGITAL TRANSITION AND ENVIRONMENTAL TRANSITION: AN UNRESOLVED CONTRADICTION

The questions of the environmental footprint of artificial intelligence and "AI for Earth" have recently been added to the agendas of decision makers with the "AI for Good" conference, in line with the United Nations's objectives for sustainable development[158], with the last World Economic Forum (2018), by the launch of the "AI for Earth" program by Microsoft (2017)[159] and with the Villani report (2018), which dedicates an entire chapter to it[160].

This placement in the agenda of a link between artificial intelligence and the environment is good news. In particular, it helps expand the discussion of potential synergies and contradictions between two great contemporary transitions: digital and environmental[161]. On the one hand, the digital transition, including megadata, artificial intelligence, the Internet of Things (IoT) and new interfaces, currently represents one of the greatest forces transforming our societies in the 21st century. On the other hand, the environmental transition is absolutely essential given three major issues: climate change, biodiversity collapse and the accelerated depletion of resources. These issues are also accompanied by serious health and social problems: strong social inequities in the face of extreme climatic events, risks to food safety in certain regions, and the impacts on health of atmospheric pollution in cities (by combustion activities that also produce greenhouse gases). They also pose a considerable challenge: Earth Overshoot Day, based on the environmental footprint concept (Rees, 1992), arrives earlier each year. The latest reports from the United Nations Environment Programme (UNEP)[162] and the Intergovernemental Panel on Climate Change (IPCC)[163] indicate that insufficient efforts are being made by countries to reduce their greenhouse gas emissions. Furthermore, the Planet Boundaries approach, which takes into consideration critical levels which, if crossed, could lead to irreversible global changes, presents a critical situation. Indeed, many limits have already been reached, and others are about to be[164].

Yet the digital transition continues to accelerate worldwide, whether for businesses (e.g. Industry 4.0), cities (smart cities) or citizens (connected mobility), with great disparity among digital consumption profiles. In 2018 the average American owned 10 connected digital devices and used 140 gigabytes of data per month, whereas the average Indian had only one and used 2 gigabytes (The Shift Project, 2018). Forecasts of acquisitions of equipment such as smartphones or the Internet of Things (IoT) by individuals and companies shows a general acceleration: by 2025, the GSMA, a telephony operator association, anticipates a net increase of 3.6 billion 4G users worldwide, and 1.2 billion new 5G users[165]. This could offer speeds of up to 10 gigabytes per second (100 times faster than 4G) and allow an intensification of mobile video use. In India, the smartphone adoption rate is expected to rise from 45% in 2017 to 74% in 2025, with 4G being the main version (62%), and the global number of connected objects should increase from 9 billion in 2017 to 55 billion in 2025[166]. This represents an explosion of data

[158] ITU (2017, 2018), *AI for Good Global Summit*, https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx and https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx

[159] Microsoft (2017), *AI for Earth can be a game-changer for our planet* https://blogs.microsoft.com/on-the-issues/2017/12/11/ai-for-earth-can-be-a-game-changer-for-our-planet/

[160] Villani C. (2018), *Donner un sens à l'intelligence artificielle,* https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

[161] Iddri, FING, WWF France, GreenIT.fr (2018), *White Paper on Digitalization and the Environment* Link: https://www.iddri.org/en/publications-and-events/report/white-paper-digital-economy-and-environment

[162] UNEP (2017), *Emissions Gap Report* Link: https://www.unenvironment.org/resources/emissions-gap-report-2017

[163] IIPCC (2018), *Special Report on Global Warming of 1.5 °C* Link: http://www.ipcc.ch/report/sr15/

[164] Earth Overshoot Day, Link: https://www.overshootday.org; Rees W. E. 1992. *Ecological footprints and appropriated carrying capacity: what urban economics leaves out*. Environment and Urbanization. 4 (2): 121-130; Rockström J. et al. 2009. *Planetary boundaries: exploring the safe operating space for humanity*. *Ecology and Society.* 14 (2): 1-33; Steffen W. et al. 2015. *Planetary boundaries: Guiding human development on a changing planet*. *Science.* 347 (6223) : 1-10.

[165] https://www.gsma.com/globalmobiletrends/

[166] https://www.businessinsider.com/internet-of-things-report

traffic on the network and in data centres. According to a Cisco report[167], worldwide traffic should increase by 25% each year (from 6.8 zettabytes in 2016 to 20.6 Zb in 2021), mainly generated by video (streaming, VOD, cloud gaming) and the Internet of Things. The storage in data centres should only increase by 36% worldwide each year (from 286 exabytes in 2016 to 1.3 Zb in 2021), the data stored on connected objects will be 5.9 Zb in 2021, 4.5 times more than that stored in data centres. The total of created (and not necessarily stored) data will reach 847 Zb per year in 2021, versus 218 Zb in 2016.

Kb, Mb, Gb, Tb, Pb, Eb, Zb … in HD movies

An HD movie consumes around 4 Gb of digital memory. Current personal computers often have a hard drive that can store 1 Tb, or about 250 movies. The Zb, which represents one billion Tb, is therefore equal to 250 billion HD movies. The total amount of data created worldwide in 2016 was equal to 218 Zb, meaning more than 7,000 movies for each person on the planet.

To communicate this data, 5G technology, with a data transfer rate of 10 Gb/s, would allow one to download the equivalent of 2 HD movies per second to a connected object.

Environmental Issues

The Shift Project[168] experts highlight that this growth can essentially be attributed to services offered by a few large companies, the American GAFAM (Google, Apple, Facebook, Amazon and Microsoft) and the Chinese BATX (Baidu, Alibaba, Tencent and Xiaomi). This growth occurs at a pace that surpasses the energy efficiency gains from the equipment, the networks and the data centres. This transition is indeed very material, and the reality of the environmental impacts, which is often swept aside or unknown, must be insisted upon.

The production of a smart phone has many impacts throughout its lifecycle, from resource extraction—issues of biodiversity, working conditions, the depletion of resources like rare earths, which incidentally are indispensable to the production of renewable energy, such as indium (used for screens and photovoltaic cells) and neodymium (used in magnets for wind turbine generators)—to the end of their lifecycle (and the problem of electronic waste, of which very little is recycled); through the use phase: energy consumption by the terminal (but also by the network and the data centre). In terms of climate change, approximately 90% of a telephone's impacts (e.g. 32 Kg CO2eq for a 5-inch phone) occur during the production period[169]. This can be explained by the fact that these phones have a very short lifespan (approx. 2 years) because of planned obsolescence. The impacts of fabrication therefore appear to be very large in a device's lifespan. GPU processors, heavily used in videogames and artificial intelligence, also consume energy[170]. Data centres also consume limited resources, such as silicium, electricity and water (for cooling). As for connected objects, they contribute to electrical and electronic waste, while consuming energy. Electronic waste is partially re-exported to developing countries where the devices are taken apart in very poor health and social conditions[171].

[167] Cisco (2018), *Cisco Global Cloud Index, Forecast and Methodology* 2016–2021, Link: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf

[168] The Shift Project (2018), *Lean ICT. Pour une sobriété numérique.* Link: https://theshiftproject.org/article/pour-une-sobriete-numerique-rapport-shift/

[169] ADEME. 2018. https://www.ademe.fr/modelisation-evaluation-impacts-environnementaux-produits-consommation-biens-dequipement and *The Schift Project* (2018), *Op. Cit.*

[170] An article in the Le Devoir newspaper (October 2018 AI series, p. 8) gives a measure of relative energy power used by the AlphaGo program and its human adversary: "In March 2016, the AlphaGo program beat the Go game champion, Lee Sedol, thanks to deep learning and learning by reinforcement, but also thanks to more than 1200 conventional processors (CPU) and at least 175 graphic processors (GPU) (…) meaning 1000 kW of power, whereas the human brain only requires 20 watts to operate." [translation]

[171] EFFACE (2015), *Illegal shipment of e-waste from the EU* (European Union action to fight environmental crime), Link: https://efface.eu/illegal-shipment-e-waste-eu-case-study-illegal-e-waste-export-eu-china; World Health Organization (2017), *Children environmental health, electronic waste*, Link: http://www.who.int/ceh/risks/ewaste/en/

The Villani report (2018)[172] quotes a report from the American association of semiconductor industrialists that predicts that in 2040, the need for storage space at the global level may exceed the available production of silicium worldwide, and that the energy required for calculation needs is also expected to exceed global energy production[173].

In the nearer term, Shift Project experts indicate that the global share of digital technologies in greenhouse gas emissions rose from 2.5% in 2013 to 3.5% in 2018, and could reach 4% by 2020 (2.1 $GtCO_2eq$). In a scenario of unchecked acceleration of the digital transition and unchanged climate policies, this would reach nearly 8% in 2025 (4.1 $GtCO_2eq$). They also indicate that the environmental footprint of digital technologies (including the energy required to build and use the equipment: servers, networks, terminals) is currently increasing by **9% each year** and captures a growing part of the world's electricity, which can compromise its decarbonation (the abandonment of fossil energy as a means to produce kWh). Lastly, they mention the likely increase of **the digital technologies' share of worldwide energy consumption.** From 1.3% in 2013, it had already doubled to 2.7% in 2017. According to their predictions, it could be anywhere from 3.2% to 6% by 2025, depending on the pace of the digital transition and the gains in energy efficiency. At 6%, the share of digital technologies would represent the consumption of over 25% more of the world's electricity in 2025!

The $GtCO_2eq$: A Measure of Greenhouse Gas Emissions

There are many types of greenhouse gases. Although carbon dioxide, or CO_2, is responsible for 76% of the global warming caused by human activity, other types must also be considered, such as methane CH4 or nitrous oxide N_2O[174]. Each gas has a different global warming potential (GWP). CO_2 is used as a reference point: its GWP is 1. Methane, for example, has a GWP of 25: one ton of CH_4 therefore has an impact 25 times greater than that of a ton of CO_2. GWP helps compare different greenhouse gas emissions, by using an equivalent ton of CO_2 (tCO_2eq) as a measuring unit.

In 2016, Canada produced 704 $MtCO_2eq$[175], the equivalent of 704 million tons of CO_2. That same year, the world produced around 50 $GtCO_2eq$.

[172] Villani C. (2018), *Donner un sens à l'intelligence artificielle*, Link: https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

[173] SIA (2015), *Rebooting the IT Revolution, a Call to Action* Link: https://eps.ieee.org/images/files/Roadmap/Rebooting-the-Revolution-SIA-SRC-09-2015.pdf

[174] Cf.: https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

[175] Cf. : https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html

Rebound effects and greenhouse gas reduction targets: the heart of a contradiction

In dynamics, this general trend can be explained by multiple rebound effects[176]. Although the energy efficiency of equipment is improving, rather than locking in these gains, we consume proportionally more goods and services: the amount of stored data increases and the devices used become more diversified (e.g. the Internet of Things), screen sizes increase, the number of potential uses continues to grow and the number of devices per user increases. Furthermore, this equipment is renewed at a very rapid pace, according to many types of obsolescence (software, algorithm, style, power, programmed). This results in an increase of greenhouse gas emissions for the sector, growing electrical and electronic waste, and pressure on rare resources and biodiversity, in particular in raw material extraction. With these rebound effects, the result is no uncoupling of digital development, on the one hand, from its materiality and environmental footprint, on the other.

These trends are in stark contrast with the greenhouse gas emission reduction objectives adopted as part of the 2015 Paris Accord to maintain global warming below 1.5 or 2 degrees compared to the preindustrial era. This contradiction increases in recent publications by UNEP[177] and IPCC[178], which indicate that an unprecedented effort to reduce our energy consumption and our greenhouse gas emissions will need to occur on a global scale within the next decade. These reports demonstrate that worldwide annual greenhouse gas emissions, which currently stand at slightly over 50 $GtCO_2eq$ per year, will need to be reduced by 10 $GtCO_2eq$ by 2030 if we are to reach the objective of 2° C, and by 20 $GtCO_2eq$ by 2030 to reach the objective of 1.5° C! And in this trajectory, which remains to be developed and exceeds existing policies and commitments made by countries, each Gigaton of CO2eq emitted annually makes a difference.

Digital Technologies Serving the Environmental Transition

Alongside the problem of the environmental footprint of digital technologies is another much more convergent perspective, through which digital applications operate as accelerators of the environmental transition (Iddri et al., 2018). In addition to smart energy networks, smart cities and smart agriculture, many innovative initiatives have found that digital technologies can be used as a participation, organization and knowledge sharing tool in the environmental transition: websites on sustainable actions or biodiversity, websites on short food circuits or ride sharing, websites on green energy co-funding, or to raise awareness about planned obsolescence, or even tele-working and videoconferencing.

Therefore, "Green IT" and "IT for Green" offer two complementary ways to think about the convergence of and contradictions between digital and environmental transitions. It is this double approach that we will adopt to discuss the relationships between artificial intelligence and the environment.

5.2

ARTIFICIAL INTELLIGENCE AND THE ENVIRONMENT: CHALLENGES AND OPPORTUNITIES

What are the specific effects of the recent boom in artificial intelligence systems (AIS), in their most recent form, machine learning, on the digital transition and the environment? We will analyze these effects by adopting two perspectives: on the one hand, the direct and indirect contributions of AIS to the environmental footprint of the digital transition, and on the other hand, the arrival of new predictive interference tools, which serve the energy and environment transition.

[176] Ray Galvin. 2015. *The ICT/electronics question: Structural change and the rebound effect.* Ecological Economics 120: 23-31.

[177] UNEP. 2017. *Emissions Gap Report 2017*. Link. https://www.unenvironment.org/resources/emissions-gap-report-2017

[178] IPCC. 2018. *Special Report on Global Warming of 1.5 °C*. Link. http://www.ipcc.ch/report/sr15/

5.2.1 Direct and indirect environmental footprint of AIS

Developing and storing databases, using sensors, developing machine learning algorithms, using new processors, developing robots equipped with AI, these are all examples of AIS. These systems represent part of the activities and technology of the digital sector, which also includes terminals such as telephones, tablets, computers, televisions, cultural activities such as videos, videogames, digital books, the Internet, and associated networks and data centres. From the viewpoint of the direct impact of their activities (energy consumption, greenhouse gas emissions, use of resources, waste and biodiversity over their lifecycle), AIS represent a part of the environmental impacts of digital technologies. Many of these points were highlighted by the participants of the deliberation and co-construction round tables organized by the Montréal Declaration for Responsible AI from February to October 2018.

However, it is in terms of their indirect effects on the global digital sector that AIS will have a major impact on the environment. Indeed, if we consider AIS and their algorithms as catalysts and accelerators in the digitization of society, with multiple rebound effects, these systems could have a critical impact on the environment. This "AI factor" in the digitization of society occurs in many ways (see the box below).

The catalyst and accelerator effect of AI on the digitization of society:

INTENSIFIED CURRENT USES: whether it's grabbing our attention through personalized recommendations, generating new images and video through GANs ("Generative adversarial networks"), augmented and virtual reality, or promises of productivity gains through Industry 4.0 or a smarter city, AI makes digital more desirable and intensifies current uses.

EXPANSION OF DIGITAL APPLICATIONS INTO NEW OBJECTS AND SERVICES: predictive services and connected personal assistants, household objects connected with vocal interaction, cobots (collaborating robots), self-driving cars with video sensors; AI allows digital technology to renew the identity of objects and services, while leading to an explosion in the data being generated, transmitted and stored.

ENVIRONMENTAL EFFECTS ON OTHER PRACTICES: personalized AI recommendations through collaborative platforms (e.g. home exchanges, purchases of secondhand goods, e-commerce) can result in environmental effects: more transportation, increased product obsolescence, etc.

ACCELERATED PACE OF EQUIPMENT RENEWAL to have **MORE POWER** and be able to use the latest artificial intelligence applications. The race to 5G for smartphones is a step in this direction, and will lead to even greater pressure on resources and the environment.

Through this structuring effect of the promotion, intensification and expansion of existing digital activities, and the accelerated pace of equipment renewal, we can expect AIS to generate much larger environmental impacts than today's digital technologies by intensifying and amplifying the rebound effects already mentioned in the previous section.

Strong sustainability

Given these changes, this document makes recommendations so that AIS and their direct or indirect environmental effects satisfy the strong sustainability requirement, compatible with the planet's environmental limits, the pace of resource and ecosystem renewal, climate stability and the non-substitutability of natural capital by artificial capital[179].

Three major solutions for strong AIS sustainability

The three solutions are as follows. The first groups information initiatives and environmental literacy on a digital platform, to allow citizens and institutional actors to have more autonomy and an improved capacity for taking initiatives. The second consists of ecodesign initiatives for companies that develop AIS. The third brings together various impactful public policies for strong AIS sustainability. In the text that follows, we describe their logic and present some inspiring examples. These solutions will be summed up in a list of recommendations in the third part of this document.

I/ INFORMATION SYSTEMS: INFORM, BUT ALSO ADVISE

Information sources on the environmental footprint of products are available with type 1 ecolabels (ISO 14,024), which guarantee that the consumer has information about the product's environmental performance over its lifecycle: the Canadian Ecologo, the European ecolabel and other ecolabels, designated type 3 (ISO 14025), more commonly used in relationships between customers and suppliers, present a summary of lifecycle analysis for the product: this is the case for the EPD (Environnemental Product Declaration), which presents a lifecycle analysis verified by a third party. Other environmental labels are used for electronic products: the IEEE1680 standard and EPEAT. Lastly, others are specifically for household appliances, which are major energy consumers (refrigerators, washing machines, etc.): the Energy Star label or the mandatory energy label on the European appliance market, which positions an appliance's energy efficiency on a performance scale in 7 to 10 classes.

Specific ecolabels that take into consideration the entire lifecycle will need to be developed for AI systems, which combine databases, sensors, interfaces, products and services into one integrated solution, and that can have indirect effects on the lifecycle (e.g. a data centre that uses kWh produced from fossil energy), as well as impacts on the digitization of society. Given the problem of planned obsolescence, which has created unprecedented pressure on resources and biodiversity, these ecolabels will also need to include criteria on extending the lifecycle of the devices used by the entire system of activities mobilized by AIS (e.g. on ecological ways to upgrade data sensors, such as user interface updates, without having to throw them away). Regarding the risk of impact related to the processing of big data, special attention needs to be paid to the data collection and storing infrastructure in the lifecycle diagnostic. An "environmental and social AIS" label will need to be developed for companies developing artificial intelligence systems for use as a selection criterion in public and private tenders, and in relationships with consumers.

Furthermore, simply informing people of the ecofriendly quality of AIS is no longer enough. Active education on the ecological use of AIS and environmental literacy about AIS must be shared, not only with citizens, but also with companies and public administrations: on planned obsolescence, capturing attention and rebound effects. For example, Iddri et al. (2018)[180] points out that tomorrow's self-driving cars, which will use AIS, could be shared in

[179] For an introduction to this concept see Bourg D. and Fragnière A. (2014), *La pensée écologique. Une anthologie*, Article: *Enjeux économiques : durabilité faible ou durabilité forte*, p. 439-443.

[180] Iddri, FING, WWF France, GreenIT.fr (2018), *White Paper on the Digital Economy and the Environment, Op. Cit.*

a public transportation mindset. But they could also remain the personal property of people who will take advantage of increased comfort to live even farther from their workplaces and turn their backs on public transportation. Another example: personalized recommendations by predictive algorithms on cultural websites try to capture the attention of users; an easy way to disconnect should always be offered, just as education on how to disconnect and be autonomous should be provided to each citizen. The way AIS is used will therefore be key to their environmental impact.

Information booklets by ADEME for the general public on the environmental issues around digital technologies provide an interesting example of this type of awareness initiative[181]. The places where such awareness-raising initiatives should be rolled out must also be carefully selected: in schools, public libraries, shops, websites using or selling AIS, etc.

Lastly, a public, free and accessible reference database on the environmental impacts of AIS and digital lifecycles should be established at the local, national and international level. The Shift Project's initiative for a Digital Environmental Directory and the ADEME's publications on the environmental impacts of consumer goods and equipment[182] are both good starting points.

II/ ECODESIGN: A CONSEQUENTIAL APPROACH FOR AIS?

For over twenty years, ecodesign initiatives, which help integrate social and environmental criteria into the product and service design and development phase[183], have made their way into many fields. In digital technologies, ecodesign initiatives and frameworks that take into account the physical lifecycle have also taken shape: *Principles for Digital Development* has a chapter entitled "Build for sustainability"[184], and a document was published on website ecodesign[185].

Given the direct and indirect environmental issues associated with AIS, it would be very useful to have an AIS ecodesign framework for companies that develop artificial intelligence solutions (e.g. a recommendation algorithm, a decision support tool, a domestic robot, a smart city system) would be very relevant. A subcommittee, ISO/IEC JTC 1/SC 42, was recently created at ISO[186] to develop an international standard framework for artificial intelligence and its ecosystem. The subcommittee could also address this question of AIS ecodesign, along with other ethical AI issues, in collaboration with the ISO/TC 207 technical committee, which is working on the ISO 14000 environmental management standards.

What are the specific issues around AIS ecodesign? How can environmental criteria be integrated into machine learning and the resulting applications? This type of work should be developed by multiparty, multidisciplinary committees. Allow us to simply highlight a few potential solutions here. The first is to adopt an approach that takes into consideration lifecycle impacts on the entire ecosystem. This approach allows an AI system to be developed and operated without causing impact transfers, like the use of equipment to collect data, data centre operations, the use of renewable energy at the highest-energy steps without diverting high-priority resources for the environmental transition, and raw material extraction and the end-of-life of equipment. The second would be to conduct a critical review of the service provided by AIS and its indirect effects to avoid environmental rebound effects (e.g. avoid capturing attention, which raises issues of user autonomy and energy overconsumption). Another path to a potential solution would be to generate a consequential lifecycle analysis initiative that would estimate the indirect environmental impacts on society associated with AIS adoption.

[181] Ademe (2017), information brochure *La face cachée du numérique*. Link: https://www.ademe.fr/face-cachee-numerique

[182] The Shift Project (2018), *Lean ICT. Pour une sobriété numérique*. Op.Cit. and ADEME (2018), Op. Cit.

[183] See for example ISO standard 14006 (2011) *Systèmes de management environnemental — Lignes directrices pour intégrer l'écoconception*. See also: Vezzoli C. and Manzini E. (2018), *Design for Environmental Sustainability.* Life Cycle Design of Products, Springer Eds.

[184] Link: https://digitalprinciples.org/principle/build-for-sustainability/

[185] F. Bordage (2015), *Eco-conception web / les 115 bonnes pratiques*, Editions Eyrolles, Paris.

[186] https://www.iso.org/committee/6794475.html

These ecodesign initiatives could be stimulated by environmental audit initiatives. The AI Now institute[187] has emphasized the importance of ethical audits for AIS in the most vulnerable sectors (education, law, health care), inspired in part by environmental law. Rather than simply operate in parallel with the environmental sector, the AI sector could also conduct audits on AIS ecodesign practises. This proposal has also been formulated by the Data and Society organization in a working paper[188]. AIS environmental evaluation platforms, such as http://www.ecoindex.fr on the environmental footprint of websites, could also be an interesting avenue.

To support these ecodesign initiatives, training programs and resources will need to be deployed: free access to quality lifecycle environmental data, public environmental databases to allow digital technology actors to analyze their environmental impact, networks to share best practices and a MOOC (Massive Open Online Course) on AIS ecodesign.

III/ PUBLIC POLICIES AND RESEARCH POLICIES: WHAT "IPCC" FOR AI?

Public policies on green and responsible procurement should be developed to systematically integrate ethical and environmental clauses into public tenders for AIS. For example, to green the value chain of AI by extending the life expectancy of equipment, banning planned obsolescence (effective in a country such as France, with its 2015 law on environmental transition) and promoting circular economic principles. Principles such as the ecodesign of data centres should also be systematically promoted by public authorities.

Furthermore, a major interdisciplinary research policy on the links between AI, digitization and environmental transition should be organized at the national and international levels. The Villani report (2018) similarly favours "establishing a space dedicated to the intersection of the environmental transition and AI" [translation]. This work could be organized in one of the current dedicated subgroups of the IPCC (International Panel on Climate Change), under its mitigation component, or in what would become a new "IPCC" on AI ethics. This research policy should cover fields of intervention as varied and important as the environmental impact of data centres (and their placement in the world to avoid diverting local resources), supply planning for rare metals in the environmental transition, electrical and electronic waste in the Internet of Things and the circular economy, the control of rebound effects and accelerating technological, software and algorithm obsolescence, the environmental benefits and ethical issues around storing DNA, machine learning with very low energy consumption, and even the emerging issues of electromagnetic smog and environmental health with the arrival of 5G in cities.

5.2.2 New predictive tools for the environmental transition

Digital technologies without AI already offer many tools that help the environment, such as a website to share environmental knowledge, a website on short food circuits, the possibility of telecommuting or taking part in a meeting without having to travel, thanks to videoconferencing, or even ride-sharing and bike-sharing platforms. In the same line of thought, AIS also offer a new range of tools for dealing with the environmental crisis. Solutions labelled "AI for Earth" have recently appeared. These rely on the specific properties of AI, such as suggesting predictive inferences in supervised learning, or classifying big data through unsupervised learning. These properties help develop tools that serve the environment:

1. a new predictive knowledge tool on social and environmental issues (e.g. on biodiversity, climate change, agricultural productivity, extreme weather events, migrations),
2. a new predictive optimization tool (e.g. for urban transportation, energy use in buildings, energy-smart grids, agriculture), and
3. a new tool to predictively regulate the environmental effects of economic actors, especially those stemming from the rebound effect.

[187] https://ainowinstitute.org/aiareport2018.pdf

[188] https://datasociety.net/blog/2018/07/03/call-for-applications-environmental-impact-of-data-driven-technologies-workshop/

Four major potential AIS solutions for the ecological transition

I/ AI AS A KNOWLEDGE TOOL SERVING THE ECOLOGICAL TRANSITION

The processing of big data by AI could help better model and understand the Earth's ecosystem. The Villani report (2018, page 127, op. cit.) presents two projects which illustrate this type of AI contribution to the environment. This includes the "Tara Oceans" project, which collects and opens big data on the ocean to better understand and model a planetary biome (ocean biodiversity and ecosystem services), and research on climate and weather, for better climate and climate risk prevention (e.g. for inhabited zones, ecosystems, agriculture).

For example, sustainable or organic agriculture can be very sensitive to extreme climate events and warming (new pests) that can cause crop failures and alter a region's food security. If AI can help improve climate forecasts and improve knowledge on resilient ecosystems, it should be used to strengthen these agricultural sustainability strategies.

II/ AI FOR EARTH TOOLBOX: BEWARE PATH DEPENDENCY

Using AI as a tool to help the environment is currently in vogue. New publications have recently presented these promising avenues in multiple ideas[189]. These suggestions are often limited to a list of very specific optimization problems (e.g. optimizing traffic flows and itineraries, smart power grids, agricultural productivity and plant protection through precision agriculture, predicting air quality), for problems sometimes inherited from former organizational, urban, agricultural and social paradigms. Although this approach has considerable potential, it must be applied rigorously to significantly contribute to sustainable development. Recent publications on AI for Earth present many shortcomings: omission of the lifecycle approach, the risks of path dependency, the rebound effects and the lack of prioritizing in regards to eco-innovation, which can cause a certain "solutionism" (the local resolution of a problem thanks to mastery of a tool, but its suboptimal use for lack of a global, integrated version). And there is no research network to critically discuss the methodology of these interventions.

In order to best use AI for the predictive optimization of polluting systems (urban transportation, energy used in building heating and cooling, agriculture, seeds and plant protection, food waste, smart energy grids, etc.), eight principles could be adopted and followed. To illustrate these principles, consider the case of an AIS project to optimize urban transportation, with a tool to make automobile traffic more fluid:

> The **lifecycle approach** (ISO 14040) to measure the impacts and benefits of these AIS and anticipate impact transfers: would the massive use of connected objects and sensors with programmed obsolescence to equip traffic lanes lead to new impacts on the lifecycle (climate change, depletion of resources, waste, biodiversity)?

> **Attention to rebound effects:** if traffic flows better and helps save time in transit, will certain users decide to live further away and therefore pollute more by contributing to urban sprawl?

> Attention to **"path dependency" mechanisms:** a bias which leads to always considering problems the same way and to optimizing the urban infrastructure with lots of available data, but with few environmental gains, while delaying a generation of sustainable breakthrough innovations (e.g. an extremely efficient and comfortable network of bike paths and public transportation).

[189] Fast (2017), *5 Ways Artificial Intelligence Can Help Save The Planet*, Link: https://www.fastcompany.com/40528469/5-ways-artificial-intelligence-can-help-save-the-planet

World Economic Forum (2018), *8 ways AI can help save the planet*, Link: https://www.weforum.org/agenda/2018/01/8-ways-ai-can-help-save-the-planet/

PwC (2018), *Fourth Industrial Revolution for the Earth. Harnessing Artificial Intelligence for the Earth*, Link: https://www.pwc.com/gx/en/sustainability/assets/ai-for-the-earth-jan-2018.pdf

- **Establishing a hierarchy of AIS according to their environmental contribution** to prioritize those that bring significant environmental benefits and avoid greenwashed "solutionism": should predictive parking, increasing the likelihood of finding a parking spot in a certain neighbourhood at a certain time, be a priority solution for the environmental transition of cities?
- The **participation** of citizens and stakeholders in the co-construction of the solutions: in the case of transportation and mobility, citizens can also help improve innovative mobility scenarios through their user experiences. A discussion on the redefinition of the desired pace of mobility in certain zones to tackle the safe coexistence of pedestrians, bicycles, self-driving cars and delivery vehicles should not only be based on past data, but also on the possibility of prospective scenarios discussed collectively.
- A **directory of AIS challenges with strong environmental potential,** to help share knowledge and experience, should be organized internationally. In our example on mobility, the C40 network of cities that have been pioneers in the fight against climate change could organize this type of community.
- **Open data policies** for public administrations as well as companies, if this data holds general interest for the environmental transition (energy, travel, biodiversity, climate, air quality, waste, etc.). This measure would help various actors develop innovative solutions to these environmental challenges, with limited data costs.
- **Digital literacy on data:** Iddri et al. (2018 op. cit.) also suggest developing a "data culture" that serves the environment through educational tools and initiatives so that all actors are able to read, create, use and communicate data, in particular public administrations and citizen groups.

III/ THE PREDICTIVE REGULATION OF REBOUND EFFECTS: POTENTIAL AND ETHICAL ISSUES

The use of AIS in the predictive algorithmic regulation of rebound effects on the consumer goods and equipment markets has considerable potential for the sustainable development of society. That would be the case, for example, of a prospective scenario where each citizen would have a three-ton carbon credit for their annual consumption, and would be encouraged not to exceed this limit through nudges and recommendations that anticipate probable rebound effects (through supervised machine learning based on past consumption behaviour data).

But this perspective raises serious ethical and democratic issues: the possible garnering of market power by a few major companies with the capacity to supply the system with certified environmental data at a lower cost than SMEs, which would be faced with a barrier to entry; the non-recognition of initiatives outside the market that nevertheless have a strong potential for the environmental transition (e.g. how can a local circular economy or sustainable mobility initiatives be valued if they are not subject to a system transaction?); the protection of privacy and the power of excessive behaviour standardization through the recommendations; the absence of a process to debate which recommendations to prioritize. Many of these points were brought up during a round table at the Montréal Declaration co-construction that focused on AIS as a tool to regulate rebound effects in society.

IV/ AI SERVING RESPONSIBLE INVESTMENT

AIS is used in market finance to equip "high-frequency trading" (HFT) devices, which are often accused of increasing the risks of a systemic financial crash, or of accelerating it, when humans lose control.

AIS could contribute to finance in other ways, by reinforcing analyses of environmental and human rights criteria for socially responsible investment. This reinforcing would occur through machine learning, like rankings in big data.

Conclusion

Given greening AIS and AIS for Earth, is it necessary to chose or prioritize one over the other to achieve strong sustainability? Given the urgent need for energy and environmental transition, both approaches should be undertaken simultaneously. The first one is needed because, due to rebound effects, there are strong unresolved contradictions between the digital and environmental transitions. The second one is required because it has significant sectoral improvement potential, as long as a certain rhetorical illusion is avoided and the principles we have presented are followed.

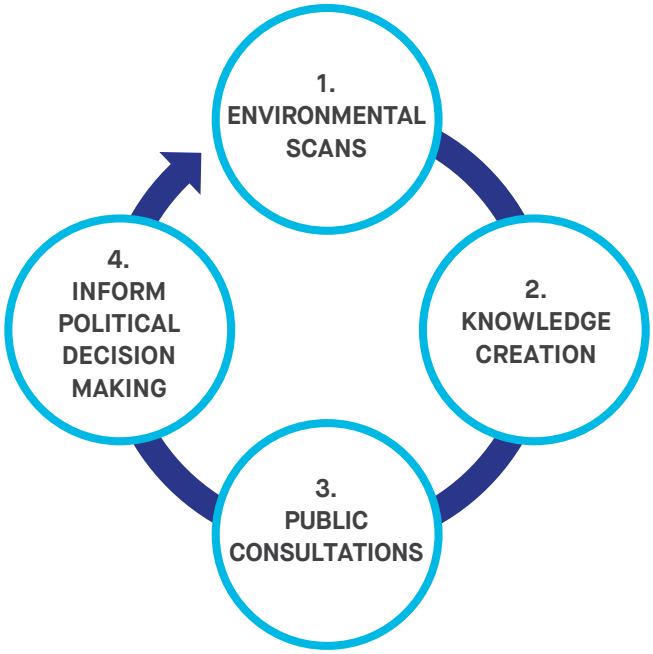
6. RECOMMENDATIONS ON THE DEVELOPMENT OF PUBLIC POLICIES

Based on the principles in the Declaration, a list of recommendations has been drawn up with the aim of suggesting guidelines for achieving the digital transition within the Declaration's ethical framework. This list should not be considered exhaustive and cannot cover all types of AI applications; nor does it include every recommendation made during the public consultations. Rather, it aims to cover a few key cross-sectoral themes for reflection on the transition towards a society in which AI is used to promote the common good: algorithmic governance, digital literacy, digital inclusion of diversity and ecological sustainability.

The recommendations that follow the Declaration are addressed more specifically to AI development actors in Quebec and Canada. They represent examples of concrete measures developed collectively from the Declaration's ethical considerations. For this reason, they can form points of convergence for actors of AI development outside Canada.

RECOMMENDATION 1: AN INDEPENDENT MONITORING AND CITIZEN CONSULTATION ORGANIZATION

We recommend establishing an organization to monitor and study the uses and social impacts of digital tools and artificial intelligence. This organization would also have a mission to help organize a participative governance space by bringing together citizens and other stakeholders to inform public policies based on environmental scans, the production of knowledge and multi-stakeholder involvement.



1.1 Establish a continuous environmental scanning mechanism that harnesses knowledge on the technical, ethical, legal and social aspects of AIS development, tracks the emergence of new issues and alerts resource persons when necessary.

1.1.1 Mobilize interdisciplinary knowledge.

1.1.2 Map best practices on algorithmic governance, with a focus on public and private partnerships and on the interests in play, the relevance of data trust models and other mechanisms associated with management of the digital commons.

1.1.3 Include citizen associations, think tanks and whistleblowers that can highlight the risks associated with AIS development.

1.1.4 Involve different types of media around digital tools and their impacts, whether it is to sound the alarm on identified relevant risks or for knowledge transfer to the general public.

1.1.5 Organize the continuous collection of feedback on the use of AIS in public and private organizations, as well as in society in general.

1.2 Foster the creation of new, diverse knowledge on the technical, ethical, legal and social aspects of AIS.

1.2.1 Conduct research on the conditions in which public automated systems can help achieve sustainable development objectives.

1.2.2 Create calls for innovative research projects, favouring inter-disciplinary approaches and a variety of viewpoints (research organizations, civil society organizations and stakeholders).

1.2.3 Produce biannual evaluation reports on the performance of public algorithms and their impacts, paying special attention to the crossover or cumulative effects of various algorithms on the situations of groups and individuals.

1.2.4 Carry out small-scale pilot projects, including within smart cities and other affected sectors, in order to determine the specific impacts of AIS in given contexts.

1.3 Mobilize citizens and stakeholders by including a proactive consultation component which will evaluate the representations and expectations of citizens as AIS develop, as their areas of activity diversify and as their reach is amplified.

1.3.1 Survey citizens on their perceptions of issues by varying survey methods (public consultations, work groups, online surveys) and by paying special attention to the socio-demographic representativeness of the participating citizens (sex, age, socio-professional environment, etc.).

1.3.2 Produce public reports that explain, in layman's terms, the results of the monitoring analysis.

1.3.3 Organize co-construction workshops that bring together citizens, civil society organizations and stakeholders to guide AIS development and rollout and make public policy recommendations.

1.4 Inform public decisions and extend the political reach of co-construction workshops through the work of experts, which consists of developing the technical aspects and recommendations, ensuring the coherence of the propositions and producing briefs and reports addressed to the policy makers and various stakeholders in AIS development.

RECOMMENDATION 2: AIS AUDIT AND CERTIFICATION POLICY

We recommend establishing a coherent AIS audit and certification policy that promotes responsible rollout (commercialization, use) of AIS and encourages stakeholders to adopt good practices to limit the adverse consequences and malicious use of AIS as much as possible.

2.1 Establish groups of multidisciplinary experts—either by using existing institutions, or by creating ad hoc groups for a limited period of time—in order to identify the institutional and legal resources that can provide potential solutions to current AI rollout issues, and identify the gaps that need to be addressed.

2.2 Extend, if required, the jurisdiction of existing institutions according to their sector and field of action (governmental associations, accreditation organizations, etc.) in order to implement an audit policy of algorithms that present a high social risk, including of human rights violations, before putting them on the market and during their use (commercial or not).

2.3 Extend, if required, the jurisdiction of existing institutions according to their sector and field of action (governmental bodies, accreditation organizations, etc.) in order to deliver AIS certifications that attest that ethical, social and legal requirements have been taken into account in AIS design, and evaluate their rollout objectives. The certification should be mandatory for all AIS used in public organizations, especially government departments.

2.4 Create a public library, accessible online, of certified AIS.

2.5 Encourage companies that develop, market or use AIS to create multidisciplinary ethics committees and internal audit process committees to identify the ethical, social and legal issues around AIS use in their commercial activities and their organization.

2.6 Develop a whistle-blowing mechanism through the creation of an online platform to gather information and complaints from individuals, groups or organizations that suspect a problem with AIS.

RECOMMENDATION 3: EMPOWERMENT

We recommend supporting citizen empowerment towards digital technologies through access to training that allows understanding, criticism, respect and responsibility that will allow citizens to actively take part in a sustainable digital society.

3.1 Promote digital literacy through a coherent education policy in primary, secondary and post-secondary establishments, to develop the skills of digital citizenship and train the next generation of scientists.

3.1.1 Integrate the teaching of digital technologies and artificial intelligence through the acquisition of fundamental technical knowledge.

3.1.2 Extend the competence of digital literacy by reinforcing the acquisition of relevant cross-disciplinary skills for full exercise of digital citizenship: using information and information technologies, exercising critical judgment, tapping into creative thinking, structuring identity, etc.

3.1.3 Reinforce the teaching of ethics regarding AI and digital issues, starting in elementary school.

3.2 Develop a policy on public spaces dedicated to digital literacy to improve access and appropriation of digital culture and encourage active citizenship and a diversity of users.

3.2.1 Offer training spaces for technological experimentation and to host digital citizen participation in third-party spaces such as public libraries, fab labs, and community and cultural centres.

3.2.2 Set aside specific funding for purchases of the necessary technological equipment and to train support staff.

3.2.3 Make training available to all through special efforts to include isolated or underrepresented groups.

> Make certain training mobile (digital knowledge trailers, mobile idea boxes).

> Prioritize specific actions targeting underrepresented groups (women, cultural minorities, etc.).

3.3 Design digital education that promotes lifestyle habits that will foster independence as well as mental and physical health throughout one's life.

3.3.1 Alert people to the risks of digital dependency, in particular by making them aware of the importance of disconnection times and spaces.

3.3.2 Support the development of non-digital skills such as pathfinding without a GPS, handwriting, etc.

3.4 Create an open-access online platform for education professionals, students, parents or tutors, and decision makers to help upgrade their knowledge on the technical, ethical, social and legal issues surrounding AI and digital technologies. In particular, this platform would be used to:

3.4.1 List organizations in the digital literacy ecosystem (educational institutions, training centres, third-party spaces, companies) and coordinate the mobilization of communities of practice in that ecosystem.

3.4.2 Guide learners, regardless of level, age or interests.

3.4.3 Establish a database of collective knowledge on AI and digital technologies.

RECOMMENDATION 4: TRAINING IN ETHICS

We recommend reviewing the training provided to those involved in the design, development and operation of AIS, making investments in multidisciplinarity and ethics.

4.1 Prioritize training for AI technicians (engineers, programmers and designers)

4.1.1 Undertake, alongside the various stakeholders, a redesign of engineering education programs to integrate knowledge on ethics, the social sciences and law so that professionals develop good intellectual reflexes, are made aware of the potentially adverse consequences of the technology they are developing, and develop creative, ethically acceptable and socially responsible solutions.

4.1.2 Promote ongoing training on social and ethics issues to ensure continued development in design and development practices and ongoing vigilance over the unexpected, undesirable effects of the AIS developed.

4.2 Extend training to workers who use AIS in the regular course of their duties and to managers who decide to adopt AIS into their organizations.

4.2.1 Ensure that the professionals using AIS understand the various aspects of their responsibility, such as being able to justify a decision made by the AIS used or based on an algorithmic recommendation, when the decision has a significant personal or social impact.

4.2.2 Ensure that they maintain their vigilance over the potentially undesirable ethical, legal and social consequences of the AIS used.

4.2.3 Make managers and social partners aware of the consequences for their organization of the digital transition, and give them the tools to carry out socially responsible restructuring.

RECOMMENDATION 5: FOSTER INCLUSIVE AI DEVELOPMENT

We recommend implementing a coherent strategy that uses the various existing institutional resources to foster inclusive AI development and prevent potential biases and discrimination related to the development and rollout of AIS.

5.1 Establish a grid of inclusion and non-discrimination technical standards for public and private AIS operations. This grid must be unique, evolving, and agreed upon by the different organizations authorized to issue regulations and professional standards (departments, professional associations). Among the provisions to be established, we recommend:

5.1.1 Testing AIS on different focus populations in order to study their impacts and uncover differences in treatment;

5.1.2 Identifying the labelling selected in the data acquisition and archiving systems (DAAS), in particular the databases used to train AIS, and the parameters guiding the decisions made by public AIS;

5.1.3 Evaluating the relevance and impact of a random parameter for ranking algorithms (search and recommendation engines), in order to reduce the importance of filtering bubbles and unavoidable biases, and ensure a diversity of recommendations that do not reflect the biases of the algorithm used;

5.1.4 Ensuring that the training databases used by public AIS contain a representative sample of the populations affected.

5.2 Integrate AIS evaluations of inclusiveness or non-discrimination performance into their certification.

5.3 Invest in programs to reinforce AI skills among groups that are traditionally underrepresented in the field, in particular women, to make their inclusion possible at every level of development, from design to application of AI technologies.

RECOMMENDATION 6: PROTECTING DEMOCRACY FROM POLITICAL MANIPULATIONS OF INFORMATION

We recommend implementing a containment strategy around information designed to trick citizens and manipulate political life on social networks and malicious web sites, as well as a strategy to fight political profiling in order to maintain conditions for healthy democratic institutions and an informed exercise of citizenship.

6.1 Organize, at different coordination levels (provincial, federal and international), a conference for stakeholders from the information and communication sector (information sites, social networks), organizations from civil society, policy makers and citizens in order to implement standards for information certification and detection of false information.

6.2 Encourage the various information sites and the press agencies that they rely on to create a joint fact-checking organization at the provincial, federal and international levels, to improve and accelerate fact-checking, to avoid a competitive verification market, to organize nonpartisan work and to increase the public's trust in information.

6.3 Promote user detection and signalling of fake news and false accounts by encouraging the common fact-checking organization, as well as web platforms (information sites, social networks), to offer their users tools that they can use to sound the alarm.

6.4 Adopt a common sign system for identifying the degree of truth in online information, on the basis of information certification standards.

6.5 Develop public AIS for detecting fraudulent sources of information on Internet platforms and encourage these platforms to develop their own detection tools.

6.6 Adopt a strategy to discourage malicious acts and slow down the propagation of false information, while avoiding situations where the measures put into place become a censoring of unpopular political opinions.

6.6.1 Systematically shut down bot accounts that spread false information.

6.6.2 Cut off advertising revenue for malicious sites and social networks that refuse to take adequate measures to prevent the spread of false information.

RECOMMENDATION 7: AI INTERNATIONAL DEVELOPMENT

We recommend adopting a non-predatory international development model aimed at including different parts of the globe without exploiting low- and middle-income countries. This model must not exploit technological backwardness or political or legal shortcomings to take their human resources (the people and data with the potential to contribute to local AI development).

7.1 Fight data appropriation by foreign companies and ensure the international traceability of data.

7.2 Ensure that the researchers, experts and decision makers from low- and middle-income countries are actively and equally involved in international discussions on AI regulation.

7.3 Support the ability of low- and middle-income countries to develop their own digital infrastructure and protect their population's data.

7.4 Create a global fund to strengthen the capacity of AI "excellence centres" in low- and middle-income countries, and invest in research programs to guide the design, development and rollout of AI.

7.5 Support international cooperation through researcher and student exchange programs between countries that are on the cutting-edge of AI development and those whose investment and development abilities are not as advanced.

RECOMMENDATION 8: DIRECT AND INDIRECT AIS ENVIRONMENTAL FOOTPRINT

We recommend implementing a public/private strategy so that the development and rollout of AIS and other digital tools is compatible with strong ecological sustainability and brings solutions to the environmental crisis.

8.1 Develop an information and awareness policy on the issues surrounding a sustainable digital transition.

8.1.1 Conduct AIS environmental audits and make them accessible so that their impact over their life cycle is known, understood and taken into consideration in purchasing and investment decisions.

8.1.2 Distribute educational information that will allow public and private organizations to steer their digital transition in a sustainable direction, paying particular attention to rebound effects and the programmed obsolence of equipment.

8.1.3 Distribute educational information that will allow citizens to adopt lifestyles leading to a very low-impact digital life.

8.1.4 Promote a techno-creative culture and foster the acquisition of skills for repairing and extending the lifespans of objects and electronics.

8.2 Develop eco-design benchmarks for AIS infrastructure and services.

8.2.1 Promote systematic AIS eco-design approaches in software development companies, accounting for their impact throughout their entire life cycle as well as the risks of rebound effects.

8.2.2 Generalize the approaches used in the eco-design of data centres and equipment (the Internet of Things, sensors and terminals using AIS) to minimize energy consumption and extend life expectancies in a circular economic logic.

8.2.3 Develop AIS and DAAS (data centres) that foster the systematic use of green electricity (renewable, decarbonated energies) at the various stages of their life cycles, without diverting this green energy from the priorities and the essential needs of local populations.

8.3 Commit to ambitious environmental public policies in response to the environmental emergency.

8.3.1 Define public policies to support research and development for digital technologies (the Internet of Things, networks, data centres, terminals) that have very low energy consumption and very small environmental footprints.

8.3.2 Implement a plan for a circular economy to reduce the need to extract the rare natural resources used by the AIS industry and better manage the flow of electrical and electronic waste.

8.3.4 Alert networks of environment and climate experts so they can develop specific knowledge on the most urgent contradictions between the ecological transition and the digital transition being accelerated by AI.

8.4 Develop and roll out AIS as a new series of tools to support the ecological transition.

8.4.1 Support the use of AIS to increase the predictive knowledge of social and environmental issues, in an open data logic, giving priority to issues surrounding climate change, the loss of biodiversity, the depletion of resources, air and water quality, in particular in major cities, and data on biomass and seeds in the context of climatic stress.

8.4.2 Support AIS development and rollout for the predictive optimization of systems with an environmental impact (initiatives called "AI for the planet") for issues such as transportation, building heating and cooling, agriculture and plant protection, the fight against food waste, and energy networks, being especially mindful of the risks of path dependency and rebound effects.

8.4.3 Experiment with using AIS as a regulation tool to predict rebound effects to establish a system that encourages sustainable consumption, compatible with respect for privacy and freedom of choice, being especially mindful of the diversity of options documented in the device.

8.4.4 Use AIS for socially responsible investment, when relevant, by calculating the carbon, social and environmental footprints of companies and institutions over their life cycles, and help make financial decisions geared towards sustainable development.

FINAL REPORT CREDITS

The Montréal Responsible AI Declaration was prepared under the direction of:

Marc-Antoine Dilhac, the project's founder and Chair of the Declaration Development Committee, Scientific Co-Director of the Co-Construction, Full Professor, Department of Philosophy, Université de Montréal, Canada Research Chair on Public Ethics and Political Theory, Chair of the Ethics and Politics Group, Centre de recherche en éthique (CRÉ)

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

Nathalie Voarino, Scientific Coordinator of the Declaration team, PhD Candidate in Bioethics, Université de Montréal

Coordination

Anne-Marie Savoie, Advisor, Vice-Rectorate of Research, Discovery, Creation and Innovation, Université de Montréal

Content contribution

Camille Vézy, PhD Candidate in Communication Studies, Université de Montréal

Revising and editing

Chantal Berthiaume, Content Manager and Editor

Anne-Marie Savoie, Advisor, Vice-Rectorate of Research, Discovery, Creation and Innovation, Université de Montréal

Joliane Grandmont-Benoit, Project Coordinator, Vice-Rectorate of Student and Academic Affairs, Université de Montréal

Translation

Rachel Anne Normand and François Girard, Linguistic Services

Rebecca Sellers, Copywriter, Translator, ESL Teacher

Graphic design

Stéphanie Hauschild, Art Director

This report would not have been possible without the input of the citizens, professionals and experts who took part in the workshops.

OUR PARTNERS

Université de Montréal

CRE
CENTRE DE RECHERCHE EN ETHIQUE

LAB VILLE PROSPECTIVE

IVADO

SAT

Mila

CIFAR
AI &
Society
Program

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