< >
Montréal Declaration
Responsible AI_
</ >

PART 6

PRIORITY PROJECTS AND THEIR RECOMMENDATIONS FOR RESPONSIBLE AI DEVELOPMENT

CREDITS

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

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

FIGURES

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.

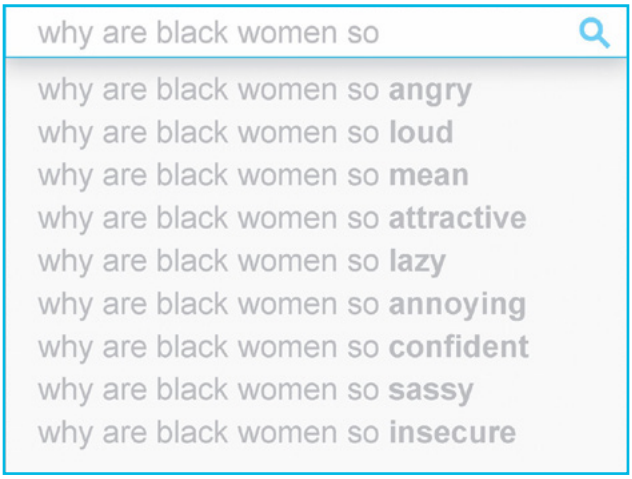
[125] See the many works of Cass Sunstein on the subject, for example: *Infotopia*, Oxford, Oxford University Press, 2006.

[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

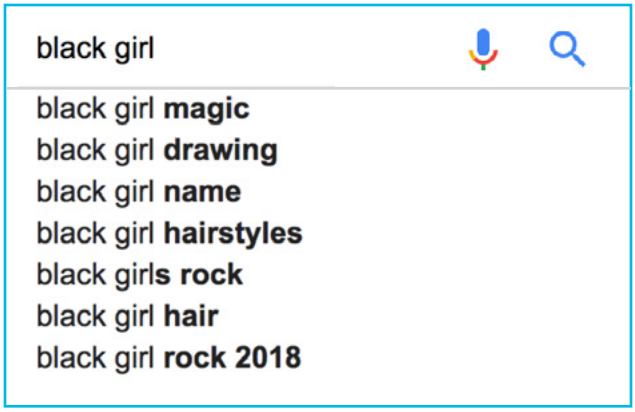
UNBIASING ARTIFICIAL INTELLIGENCE SYSTEMS

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.

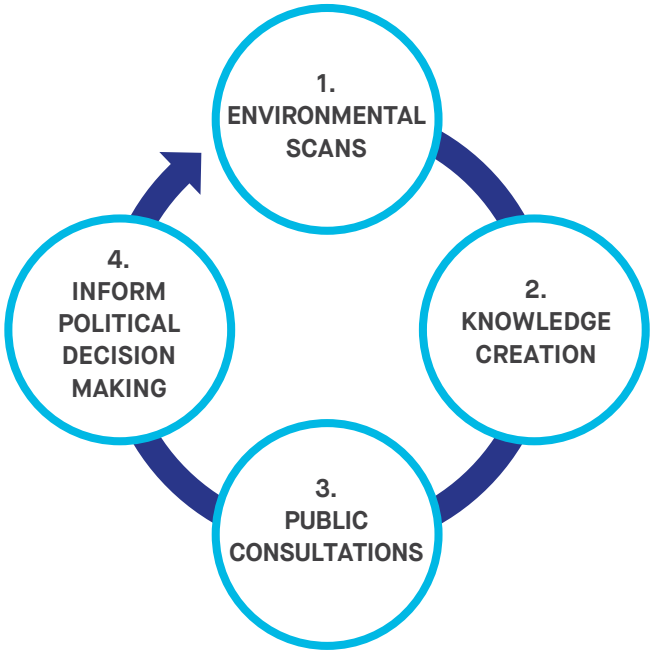
6. RECOMMENDATIONS ON THE DEVELOPMENT OF PUBLIC POLICIES

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