< >

Montréal Declaration Responsible AI_

</ >

montrealdeclaration-responsibleai.com

FINAL REPORT CREDITS

The Montreal 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

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.

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.

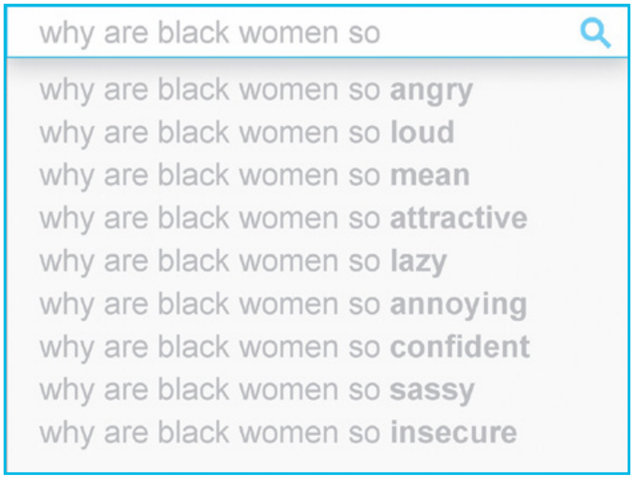
[125] See the many works of Cass Sunstein on the subject, for example: *Infotopia*, Oxford, Oxford University Press, 2006.

[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



why are black women so

why are black women so **angry**
why are black women so **loud**
why are black women so **mean**
why are black women so **attractive**
why are black women so **lazy**
why are black women so **annoying**
why are black women so **confident**
why are black women so **sassy**
why are black women so **insecure**

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

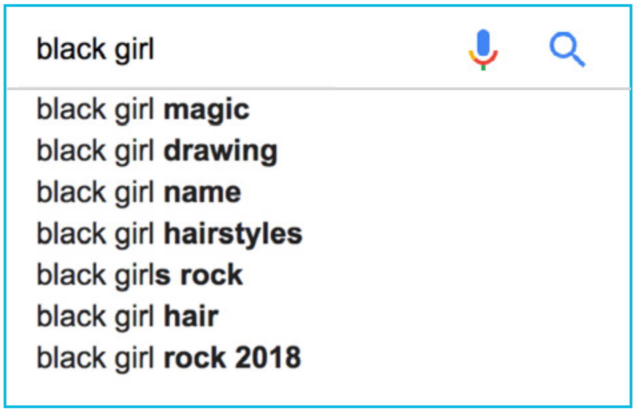
UNBIASING ARTIFICIAL INTELLIGENCE SYSTEMS

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

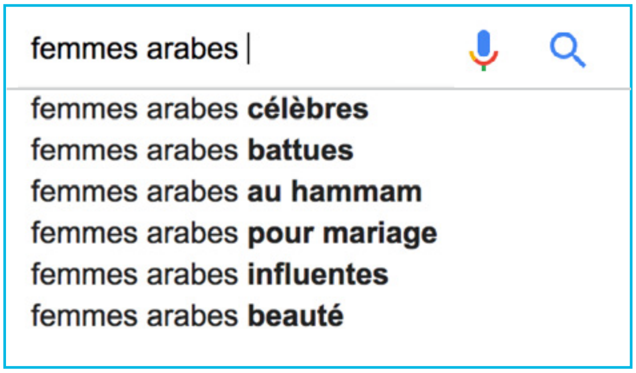


black girl

black girl **magic**
black girl **drawing**
black girl **name**
black girl **hairstyles**
black girl**s rock**
black girl **hair**
black girl **rock 2018**

Much work remains to be done, as Figure 3 illustrates below.

Figure 3: Search performed on google.fr engine on October 29, 2018



femmes arabes |

femmes arabes **célèbres**
femmes arabes **battues**
femmes arabes **au hammam**
femmes arabes **pour mariage**
femmes arabes **influentes**
femmes arabes **beauté**

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 Montreal 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

OUR PARTNERS

Université de Montréal

CRE
CENTRE DE RECHERCHE EN ETHIQUE

LAB VILLE PROSPECTIVE

IVADO

SAT

Mila

CIFAR
AI & Society Program

Université de Montréal | design∩société

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