Montreal Declaration for Responsible AI

Proposed Edits from the Montreal AI Ethics Meetup community

https://www.meetup.com/Artificial-Intelligence-Ethics/

Submission prepared by:

Abhishek Gupta - Founder, Montreal AI Ethics Institute, AI Ethics Researcher, McGill University, Software Engineer - Machine Learning, Microsoft

Stephanie Dyke - PhD, Montreal Neurological Institute, McGill University

Paule-J Toussaint

Gregory Caicos PhD

Marc Daher - B.Eng. Software & Multimedia

Peter Chen

Background to this proposal:

1. The comments submitted in this document are a summary of discussions involving over a 100 multidisciplinary scholars, including AI researchers, and concerned citizens, who met during eight 2-hour bi-monthly AI Ethics Meetup sessions to review and discuss the principles of the Montreal Declaration.
2. In general, this was a great exercise for the local Montreal AI ecosystem. We also believe the set of principles, along with points-to-consider for further guidance, are an important contribution to international discussions in this area, and indeed complementary to the recent IEEE Ethically Aligned Design, Asilomar Principles and UK House of Lords Select Committee recommendations.
3. If we wish to make a “distinct” Montreal contribution, then, we should consider further what exactly would be our *special* contribution, its goals and the community(ies) we wish to represent, so that we are clearly acting towards a common cause and contributing productively to international efforts.
4. What follow are the thoughts summarized and organized by suggested reframing of the principles and points to consider that would further enrich the declaration.

Privacy

> *The development of AI should offer guarantees respecting personal privacy and allowing people who use it to access their personal data as well as the kinds of information that any algorithm might use.*

We have serious concerns with the current drafting of this principle and propose the following text:

> *The development of AI should protect privacy and be governed to enable oversight of, and consent to, access to personal data and their use in every AI application. Transparency, guidelines and regulation will be essential to this task.*

We strongly feel that governance and regulation of AI will be essential to preserve privacy rights and respect consent to the use of personal data in AI development and applications. Privacy rights encompass both group (societal or collective), as well as more "individual", forms of privacy, and privacy preferences are known to vary considerably between cultures, generations, and individuals.

We propose the following points-to-consider for further development of this principle:

Governance & Regulation

1. Monitoring and audit to check data uses to help with compliance
2. Data audit requirements, processes, methods, standards, and code of conducts
3. Reporting standards linked to monitoring and audits
4. Transparency will be key to oversight
5. Public & private sector regulation
6. Tie in with privacy/data protection laws
7. Do we have the institutions, tools, and capacity to regulate this?

Consent & Withdrawal

1. People need to be informed about, and allowed and able to check, their personal data and its uses at any time
2. Individuals must understand which of their personal data is being used and how it is being used in AI development and applications
3. What are the barriers to free and informed consent to the use of personal data (illiteracy, legal [il]literacy, complexity)?
4. A reasonable level of choice must be provided for in consent processes (not all or nothing situations), with the possibility to withdraw consent at any time
5. What do AI black boxes mean for informed consent?
6. The wording of "consent forms" and terms & conditions of use is incomprehensible to most. Should provide means for vulnerable groups, including children, to understand these terms and conditions.

Security

1. What level of security is needed to protect personal data in their use in AI development and AI applications?
2. Consider levels of data protection based on data sensitivity

Social media

1. Consent & transparency are key
2. Should provide protection / protected areas / buffered areas for vulnerable groups (fragile psyche, children).
3. Should provide / offer help - easily accessible (button) so that vulnerable groups, including children, could seek help in situations of distress and feel / be supported.
4. Need for (free) education / instructions on usage of social media, and how to be a good netizen (recognising / stopping abuse, limits of what can be said / written and published / uploaded on social media).

Justice

> *The development of AI should promote justice and seek to eliminate all types of discrimination, notably those linked to gender, age, mental / physical abilities, sexual orientation, ethnic / social origins and religious beliefs.*

We have some concerns with the way the principle is currently framed and propose the following text for this value:

> *The development and utilization of AI-enabled solutions should promote justice and human agency as transparently defined by the target community's welfare-defining organization (e.g. democratically elected government), in concert with the target community. It should seek to eliminate inequality and discrimination within that community.*

We believe that this framing allows for a more comprehensive definition while emphasizing contextual and cultural differences between different target audiences that are going to be the subject of AI-enabled solutions under consideration. Additionally, when defining the norms around justice, equality and other related objectives and constraints for the system, there should be an inclusive stakeholder consultation that will allow the voices of the governed to be an integral part of decisions on how they are governed. Given that humans embody notions of empathy and common-sense that are the best at the moment, they must be an integral part of justice decisions. The system should be setup in a way such that it can be amended as society evolves (taking hints from the way a constitution allows for amendments)

Some of the other points to consider as this value of Justice is developed further:

Preventing discrimination and removing bias:

1. We need to consider the different perspectives within target communities, especially those of marginalized communities, in defining the notions of justice, equality and discrimination.
2. Cultural and contextual values need to be at the center of defining these notions and ideally should be in tight consultation with representative members from the target community.

Inequality

1. What mechanisms can we put in place to counter the concentration of wealth and power in the hands of a few such that they can distort economical, political and societal institutions?
2. How do we make sure that the gains from developing and utilizing an AI-enabled solution are equitably distributed, ideally distributed so that they benefit marginalized communities?
3. If there comes a divide between humans *augmented* by the use of such AI-enabled solutions while there are those that aren't, how do we manage the potential resulting inequalities?
4. Education around the development and utilization of these solutions must be a priority to minimize the gaps between abilities, distribution of gains, etc. within the target community

Transparency

1. When taking decisions regarding any of the above, the process should be transparent and open-source allowing for feedback from as many participants within that target community as possible.
2. The regulations and guidelines around this value should be interpretable to the common man in a way that they can meaningfully exercise their rights
3. We should consider evaluation of the transparency and interpretability of the above process by a third-party to ensure that we meet the requirements allowing for inclusive and representative participation

Agency

1. The solutions should be such that they allow for the maximization of human choices while not infringing on the rights of other humans
2. There should be strong due process to contest decisions, even when a human was a part of the decision rendered by an AI-enabled solution.
3. Should the AI-enabled solution be allowed to pursue notions of justice, equality, and non-discrimination at a collective level if it renders harm at an individual level?

Knowledge

The development of AI should promote critical thinking and protect us from propaganda and manipulation.

Our major concern with this principle as currently drafted is its statement about AI protection from propaganda and manipulation. Although we recognize there may be AI tools, such as "fact-checking" tools, that could help assess the quality of information, AI and its producers or suppliers are very unlikely to be in a position to judge, or indeed protect anyone from,

propaganda and manipulation. We feel a statement about openness and transparency would be much more impactful in this regard, allowing for public participation in, and scrutiny of, AI development. This would also further the goals of public education in AI which will play a critical role in protecting the public from abuses of AI technologies.

We therefore propose the following text for the 'Knowledge' principle:

> *The development of AI should not hamper critical thinking. It must also proceed in a transparent and open manner, to enable public participation in its development, scrutiny, and education. In particular, measures should be in place to promote public access to academic AI research results.*

We propose the following points-to-consider to further develop this principle:

Public access to AI research

1. Data and source code sharing policies must be strengthened along with open access publication policies.
2. There are particular reproducibility difficulties in AI research, which need to be addressed by research communities.
3. Access to AI technology should benefit society through greater competition and diversity in AI applications and solutions.

Business incentives

1. Raise awareness of business incentives that inevitably lead to "echo-chamber" effects.
2. Rethink business models for social media and other social news sites.
3. Companies are unlikely to share AI algorithms due to IP interests so transparency and responsible development will require regulation.

AI and critical thinking

1. It could be possible that AI might hamper critical thinking by reducing certain mental faculties (e.g. google maps replacing orientation), and have negative effects on other areas of function, such as relationships. Research into such impacts would be beneficial.

Democracy

(The name of the principle itself can be reframed as Public Participation)

Proposed principle:

> *The development of AI should promote informed participation in public life, cooperation and democratic debate.*

We have concerns with the proposed drafting of this principle, in particular with the vagueness of the wording and the undercurrent ideas of controlled social interactions and pertinence of areas of research. We believe that democracy involves constant public education, in which case it also entails transparency and the sharing of information about AI research, be it technical knowledge or a more general description of the scientific or social framework.

We propose the following text for the 'Democracy' principle, to be renamed 'Public Participation':

> *The development of AI should promote the dissemination of clear and accurate information to the public to enable open and educated debate about AI and its applications, and encourage open and transparent research collaboration.*

The proposed changes are further elucidated in the three points-to-consider below.

Democratic debate:
There should be an educated debate rather than a democratic one given that the world governments are not all based on the notion of Democracy. We propose renaming this principle 'Public Participation', and using the term "educated debate" to bring forth the notions of education and public guidance.

Cooperation
1. Open datasets and code: both are an integral part of the development of AI research, and their accessibility would allow for democratisation of research.
2. Open research and publishing: these depend on point 1, and in certain scientific communities (e.g. neurosciences) there are task forces currently at work to establish publication guidelines and standards for data and code.
3. Open collaboration: scientific advances and collaboration should transcend geographical and political barriers, and promote a healthy competition while putting forward human autonomy and well-being.

The above three points are also firmly tied to the "Knowledge" value.

Informed participation in AI development
Keeping the public informed on the subject AI research will require transparency and clear communication of current AI research status and progress, and intelligible debate of how it could / will affect their everyday lives and social interactions.

Well-being

Proposed principle:

> *The development of AI should ultimately promote the well-being of all sentient creatures.*

Our proposed edit:

> *The development of AI should aim to alleviate human suffering and bring about the well-being of all sentient life, while maintaining human freedom of choice.*

Reasons for the rewording:

Alleviate human suffering: traditionally, preventing harm as a concept has more easily led to concrete action than promoting well-being. Further, the focus on human suffering centres the principle on humanity's well-being, while not disregarding all sentient life.

"Promote": This word does not indicate strongly enough that AI must bring about well-being. That is, AI could be a toxic factory, that happens to have billboards up that promote the good life. We suggest changing this to indicate that AI will prioritize, and be intrinsically involved in the prevention of harm and the causing of well-being, as well as the promotion of it. This would be consistent with the general call for beneficial AI.

"Creatures": This word indicates a Judeo-Christian bias from its root "creation", indicating that life comes from a Creator-God. Instead, we propose the replacement of "creatures" with "life". Both "sentient" and "life", placed together, imply, at least in English, a care for the biosphere, while implicitly acknowledging that there is a gradation of sentience and complexity in life.

We stress that maintaining human freedom of choice is crucial here so as to avoid situations in which the goal of well-being could be used to justify unnecessary constraints on human freedom, including those that may impact on the Autonomy principle.

The principle could benefit from further specificity, given the contemporary critical debates on the definitions of well-being and sentience. But, as a principle it needs to be simple, clear and understandable. We acknowledge that any apparent vagueness here, however minimal, may be a good thing in this case, in that it invites future interpretation without locking the concepts down to our own biases.

Autonomy

Proposed statement:

> *The development of AI should promote the autonomy of all human beings and control, in a responsible way, the autonomy of computer systems.*

We have some concerns with the way the principle is currently framed and propose the following text for this value:

> *The development of AI should respect and promote the autonomy of all humans and enhance their self-determination while not hindering the growth of wellbeing, public participation, knowledge, responsibility, justice and privacy.*

We believe that this framing allows for the important notion of self-determination to be highlighted in the wording of the principle which is crucial to the notion of autonomy of human beings.

In addition, the removal of the phrase “the autonomy of computer systems” keeps in mind some of the intended/unintended connotations of letting the autonomy of the computer systems get in the way of *eudaemonia* at the highest level and also potentially the hindering of successful implementation of other principles.

Ideas of run-away objective functions, due to corrupted reward channels among other things, have shown that a higher degree of autonomy can, at times, lead to the system arriving at pathways to achieve objectives that achieve the objective but fail to keep with the “spirit” of the objective. https://arxiv.org/pdf/1803.03453.pdf

As highlighted with other principles in this proposal, the development of AI systems should be “sentient life”-centric, rather than pushing for the autonomy of the systems for the sake of autonomy, except in cases where the increased autonomy of the computer systems themselves leads to better outcomes for sentient life and are in line with implementations of the other principles identified in this proposal.

Consent is the primary mechanism through which individuals are able to express their autonomy, e.g., to the use of an AI or of their personal data.

Responsibility

Proposed principle:

> *The various players in the development of AI should assume their responsibility by working against the risks arising from their technological innovations.*

Our proposed edit:

> *AI designers and their sponsors (academic, military or corporate) should assume full responsibility, accountability and liability for any negligent, unjust or catastrophic outcomes facilitated by any AI they produce. This includes building the technical knowledge and competence to understand the workings and anticipate the reactions of AI.*

Reasons for the rewording:

"Players": This term is not a exact cognate of an "actor" in French in this context. Both actor and player can indicate mostly bad/shifty people, rather than including good, well-meaning people.

Whose Moral Agency?: AI developers and society should avoid granting AI's moral agency or patiency (as argued by Bryson, 2016 and Johnson, 2006). An AI should be considered an extension of, but not autonomous to, human intentionality. To that end, AI should not have a deceptive appearance that would entice humans to grant them empathy-deserving moral patiency (like a plush toy). As well, their workings need to be transparent, as argued elsewhere here. We recognize that the main proponents of those wishing to grant an AI moral agency are those most likely to benefit: the advertisers, designers and corporations who wish to evade responsibility for their designs, and profit from them.

"Working against the risks": An "intention" to work against risk is not enough in our opinion. This says little and is confusing. A thief works against the risk of being caught. A well-meaning, but distracted gas-station employee, working against risk, can still set a neighbourhood ablaze. AI developers must assume responsibility and liability to avoid negligence, unjust actions and dangerous outcomes. This principle, in light of recent Facebook privacy breaches, self-driving cars killing pedestrians, and autonomous weapons, we think, needs to have teeth.

We have also added to the principle that designers and developers of AI should demonstrate that they have the technical know-how and competence to avoid or anticipate and correct "black box" situations.

Regulation will be essential to defining responsibilities and delimiting responsible development of AI for all parties involved.