**ARCHIVE : Version préliminaire de la :**

**Déclaration de Montréal**

**pour un développement responsable de l’intelligence artificielle.**

**(Phase 1)** Version du 23/10/17

**Préambule**

Qu’elle soit naturelle ou artificielle, l’intelligence n’a pas de *valeur en soi*. L’intelligence d’un individu ne nous dit rien de sa valeur morale; c’est aussi le cas pour toute autre entité intelligente. L’intelligence peut néanmoins avoir une *valeur instrumentale*: c’est un outil qui peut nous éloigner ou nous rapprocher d’un objectif que nous valorisons. Ainsi, l’intelligence artificielle (IA) peut créer de nouveaux risques et exacerber les inégalités économiques et sociales. Mais elle peut aussi contribuer au bien-être, à la liberté ou à la justice.

D’un point de vue éthique, le développement de l’IA pose des défis inédits. Pour la première fois dans l’histoire, nous avons la possibilité de créer des agents non humains, autonomes et intelligents, qui n’ont pas besoin de leurs concepteurs pour accomplir des tâches que l’on croyait réservées à l’esprit humain. Ces machines intelligentes ne se contentent pas de mieux calculer que les êtres humains, elles cherchent, traitent et diffusent des informations. Elles interagissent avec des êtres sensibles, humains ou non humains. Bientôt, elles pourront même leur tenir compagnie, comme un parent ou un ami.

Ces agents artificiels vont donc être amenés à influencer directement nos vies. À long terme, on pourrait créer des « machines éthiques », c’est-à-dire capables de prendre des décisions en se conformant à des principes éthiques. Il faut se demander si ces développements sont responsables et souhaitables. Et il est permis d’espérer que l’IA rende nos sociétés meilleures, dans l’intérêt de tous et le respect de chacun.

Les principes et les recommandations que nous vous invitons à élaborer collectivement sont des orientations éthiques pour le développement de l’intelligence artificielle. Pour cette première phase de la déclaration, nous avons identifié sept valeurs: bien-être, autonomie, justice, vie privée, connaissance, démocratie et responsabilité. Pour chacune d’entre elle, vous trouverez une série de questions qui visent à explorer sa relation avec le développement de l’IA. Nous proposons ensuite un principe général - qui ne répond pas directement aux questions soulevées.

**1- Bien-être.**

* Comment l’IA peut-elle contribuer au bien-être ?
* Est-il acceptable qu’une arme autonome puisse tuer un être humain ? Un animal?
* Est-il acceptable qu’une IA contrôle un abattoir ?
* Devrait-on confier la gestion d’un lac, d’une forêt ou de l’atmosphère terrestre à une IA?
* Devrait-on développer des IA capables de ressentir du bien-être ?

*Principe proposé :* ***Le développement de l’IA devrait ultimement viser le bien-être de tous les êtres sentients.***

**2- Autonomie**

* Comment l’IA peut-elle contribuer à l’autonomie des êtres humains?
* Faut-il lutter contre le phénomène de capture de l’attention dont s’accompagnent les avancées de l’IA ?
* Faut-il s’inquiéter de ce que des humains préfèrent la compagnie des IA à celle d’autres humains ou d’animaux ?
* Peut-on donner son consentement éclairé face à des technologies autonomes de plus en plus complexes?
* Faut-il limiter l’autonomie des systèmes informatiques intelligents? Un humain devrait-il toujours avoir la décision finale?

*Principe proposé :* ***Le développement de l’IA devrait favoriser l’autonomie de tous les êtres humains et contrôler, de manière responsable, celle des systèmes informatiques.***

**3- Justice**

* Comment s’assurer que les bénéfices de l’IA soient accessibles à toutes et à tous?
* Faut-il lutter contre la concentration du pouvoir et de la richesse au sein d’un petit nombre d’entreprises en IA?
* Quelles sont les discriminations que l’IA pourrait créer ou exacerber ?
* Le développement de l’IA devrait-il être neutre ou chercher à réduire les inégalités économiques et sociales?
* Quels types de décisions de justice pourrait-on déléguer à une IA?

*Principe proposé :* ***Le développement de l’IA devrait promouvoir la justice et viser à éliminer les discriminations, notamment celles liées au genre, à l’âge, aux capacités mentales et physiques, à l’orientation sexuelle, aux origines ethniques et sociales et aux croyances religieuses.***

**4- Vie privée**

* Comment l’IA peut-elle garantir le respect de la vie privée ?
* Nos données personnelles nous appartiennent-elles et devrait-on avoir le droit de les effacer?
* Devrait-on savoir à qui nos données personnelles sont transmises et, plus généralement, qui les utilise ?
* Est-il contraire aux règles d’éthique ou d’étiquette qu’une IA réponde à votre place à vos courriels ?
* Qu’est-ce qu’une IA pourrait faire en votre nom?

*Principe proposé :* ***Le développement de l’IA devrait offrir des garanties sur le respect de la vie privée et permettre aux personnes qui l’utilisent d’accéder à leurs données personnelles ainsi qu’aux types d’informations que mobilise un algorithme.***

**5- Connaissance**

* Le développement de l’IA fait-il courir un risque à la pensée critique?
* Comment minimiser la circulation de fausses nouvelles ou d’informations mensongères?
* Les résultats des recherches (positif ou négatifs) en IA doivent-ils être disponibles et accessibles?
* Est-il acceptable de ne pas être informé que des conseils médicaux ou légaux sont donnés par un chatbot?
* En quel sens les algorithmes devraient-ils être transparents quant à leur processus de décision?

*Principe proposé :* ***Le développement de l’IA devrait promouvoir la pensée critique et nous prémunir contre la propagande et la manipulation.***

**6- Démocratie**

* Faut-il contrôler institutionnellement la recherche et les applications de l’IA?
* Dans quels domaines est-ce le plus pertinent?
* Qui devrait décider - et selon quelles modalités - des normes et valeurs morales déterminant ce contrôle ?
* Qui devrait choisir les « réglages moraux » des voitures autonomes ?
* Faudrait-il développer un ou des labels « éthiques » pour les IA, les site web ou les entreprises qui respectent certains standards ?

*Principe proposé :* ***Le développement de l’IA devrait favoriser la participation éclairée à la vie publique, la coopération et le débat démocratique.***

**7- Responsabilité**

* Qui sont les acteurs responsables des conséquences du développement de l’IA?
* Comment définir un développement progressiste ou conservateur de l’IA ?
* Comment réagir devant les conséquences prévisibles sur le marché du travail?
* Est-il acceptable de confier une personne vulnérable aux bons soins d’une IA ? (par exemple, avec un « robot-nanny »)
* Un agent artificiel comme Tay, le chatbot « raciste » de Microsoft, peut-il être moralement blâmable et responsable?

*Principe proposé :* ***Les différents acteurs du développement de l’IA devraient assumer leur responsabilité en œuvrant contre les risques de ces innovations technologiques.***

**LEXIQUE**

**Êtres sentients:** c’est un être capable d’éprouver du plaisir, de la douleur, des émotions et, plus généralement de *ressentir*. Dans l’état actuel des connaissances scientifiques, on peut dire que l’ensemble des vertébrés et certains invertébrés comme les poulpes sont des êtres sentients. En biologie, le développement de la sentience peut s’expliquer par la théorie de l’évolution.

**Éthique (ou morale)**: c’est la discipline qui réfléchit à la bonne manière d’agir, individuellement ou collectivement, en cherchant à adopter un point de vue impartial. Elle s’appuie sur des normes et des valeurs morales.

**Valeurs morales**: les valeurs morales sont relatives au bien et au mal: elles permettent, par exemple, de qualifier une action comme juste ou injuste, honnêtes ou malhonnêtes, louable ou blâmable.

**Valeurs épistémiques**: les valeurs épistémiques sont relatives à la connaissance: elles permettent, par exemple, de qualifier un argument comme bon ou mauvais, claire ou fumeux, pertinent ou futile.

**Valeur intrinsèque**: une valeur est intrinsèque lorsqu’elle est une justification ultime, lorsqu’on peut la rechercher *pour elle-même*. Par exemple, le bien-être, l’autonomie ou la justice peuvent être recherchés pour eux-mêmes: ce sont donc des valeurs intrinsèques.

**Valeur instrumentale:** une valeur est instrumentale lorsqu’elle est au service d’autre chose, lorsqu’elle permet de promouvoir une valeur intrinsèque, par exemple. Ainsi l’argent ou l’intelligence sont des valeurs instrumentales qui peuvent être mise au service du bien-être, de l’autonomie ou de la justice.

**Utopie:** c’est un monde possible où s’incarne un ensemble de valeurs positives. Ainsi, on peut dire qu’une société où l’IA aurait libérée les gens de tous les travaux pénibles tandis qu’ils prendraient soins les uns des autres tout en développant leurs potentiels personnels, est une société utopique.

**Dystopie**: c’est le contraire d’une utopie, un monde possible où s’incarne un ensemble de valeurs négatives. Ainsi, on peut dire qu’une société où quelques (ou une seule) entreprises devenues hyper-puissantes grâce à l’IA contrôlerait et exploiterait la population, est une société dystopique.

**RECOMMANDATIONS** (à venir) :

**Montreal Declaration for the Responsible Development of Artificial Intelligence.**

**(Phase 1)**

**Preamble**

Intelligence, whether it be natural or artificial, has no value in and of itself. An individual’s intelligence does not tell us anything about his / her morality; this is also the case for any other intelligent entity. Intelligence can, however, have an *instrumental value*: it is a tool that can lead us away from or towards a goal we wish to attain. Thus, artificial intelligence (AI) can create new risks and exacerbate social and economic inequalities. But it can also contribute to well-being, freedom and justice.

From an ethical point of view, the development of AI poses previously unknown challenges. For the first time in history, we have the opportunity to create non-human, autonomous and intelligent agents that do not need their creators to accomplish tasks that were previously reserved for the human mind. These intelligent machines do not merely calculate better than human beings, they also look for, process and disseminate information. They interact with sentient beings, human or non-human. Soon, they will even be able to keep them company, as would a parent or a friend.

These artificial agents will come to directly influence our lives. In the long term, we could create “moral machines”, machines able to make decisions according to ethical principles. We must ask ourselves if these developments are responsible and desired. And we can hope that AI will make our societies better, in the best interest of, and with respect for, everyone.

The principles and recommendations that we are asking you to develop collectively are ethical guidelines for the development of artificial intelligence. In this first phase of the declaration, we have identified seven values: well-being, autonomy, justice, personal privacy, knowledge, democracy and responsibility. For each value, you will find a series of questions that seek to explore its relationship with the development of AI. We then put forth a general principle, but one that does not directly answer the questions asked.

**1 – Well-Being**

* How can AI contribute to personal well-being?
* Is it acceptable for an autonomous weapon to kill a human being? An animal?
* Is it acceptable for AI to control an abattoir?
* Should we entrust AI with the management of a lake, a forest or Earth’s atmosphere?
* Should we develop AI which is able to sense well-being?

*Proposed Principle*: **The development of AI should ultimately promote the well-being of all sentient beings.**

**2 - Autonomy**

* How can AI contribute to greater autonomy for human beings?
* Must we fight against the phenomenon of attention seeking which has accompanied advances in AI?
* Should we be worried that humans prefer the company of AI to that of other humans or animals?
* Can someone give informed consent when faced with ever more complex autonomous technologies?
* Must we limit the autonomy of intelligent computer systems? Should a human always make the final decision?

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

**3 - Justice**

* How to ensure that the benefits of AI are available to everyone?
* Must we fight against the concentration of power and wealth in the hands of a small number of AI companies?
* What types of discrimination could AI create or exacerbate?
* Should the development of AI be neutral or should it seek to reduce social and economic inequalities?
* What types of legal decisions can we delegate to AI?

*Proposed Principle*: **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.**

**4 – Personal Privacy**

* How can AI guarantee respect for personal privacy ?
* Do our personal data belong to us and should we have the right to delete them?
* Should we know with whom our personal data are shared and, more generally, who is using these data?
* Does it contravene ethical guideline or etiquette for AI to answer our e-mails for us?
* What could AI do in your name?

*Proposed Principle*: **The development of AI should offer guarantees concerning the respect of personal privacy, better control of personal data, and access to any kinds of information that an algorithm might use.**

**5 – Knowledge**

* Does the development of AI put critical thinking at risk?
* How to minimise the dissemination of fake news or misleading information?
* Should research results on AI, whether positive or negative, be made available and accessible?
* Is it acceptable not to be informed that medical or legal advice has been given by a chatbot?
* In what ways should algorithms be transparent as to their internal decision making processes?

*Proposed Principle*: **The development of AI should promote critical thinking and protect us from propaganda and manipulation.**

**6 - Democracy**

* Must AI research and its applications, at the institutional level, be controlled?
* In what areas is this the most pertinent?
* Who should decide, and according to what modalities, the norms and moral values determining this control?
* Who should choose the “ethical guidelines” for self-driving cars?
* Must one or several “ethical labels”, which respect certain standards, be developed for AI, web sites or businesses?

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

**7 - Responsibility**

* Who is responsible for the consequences of the development of AI?
* How to define progressive or conservative development of AI?
* How to react when faced with AI’s predictable consequences on the labour market?
* Is it acceptable to entrust a vulnerable person to the care of AI? (for example, with a “robot-nanny”.)
* Can an artificial agent, such as Tay, Microsoft’s “racist” chatbot, be morally culpable and responsible?

*Proposed Principle*: **The different actors in the development of AI should assume their responsibility by working against the risks arising from these technological innovations.**

**Lexicon :**

**Sentient Being :** Any being able to feel pleasure, pain, emotions; basically *to feel*. In the current state of scientific knowledge, all vertebrates and some invertebrates, such as octopi, are considered sentient beings. In biology, the development of this characteristic can be explained by the theory of Evolution.

**Ethics (or Morals)** : This is the discipline that ponders the proper ways to behave, individually or collectively, by looking to adopt an impartial point of view. It is based on moral norms and values.

**Moral Values:** Moral values are related to good and evil: they allow us, for example, to qualify an action as just or unjust, honest or dishonest, commendable or blameworthy.

**Epistemic Value**: Epistemic values are related to knowledge: they allow us, for example, to qualify an argument as valid or invalid, clear or unclear, pertinent or trivial.

**Intrinsic Value :** A value is intrinsic when it is an ultimate justification, when one looks for it *in and of itself*. For example, well-being, autonomy or justice can be looked in and of itself : thus they are intrinsic values.

**Instrumental Value :** A value is instrumental when it is in service of something else, when it helps promote an intrinsic value, for example. Money and intelligence are such instrumental values that can be put to the service of well-being, autonomy or justice.

**Utopia :** A possible world which embodies a collection of positive values. Thus, it can be said that a society in which AI frees people from all unpleasant work, allowing them to take care of each other while fully developing their personal potential, would be a utopian society.

**Dystopia :** This is the opposite of a utopia. It is a possible world that embodies a collection of negative values. Thus, it can be said that a society where several corporations (or a single corporation) become extremely powerful thanks to AI allowing them to control and exploit the population, would be a dystopian society.

**RECOMMANDATIONS** (to come) :