ÉLÉMENT DE PORTFOLIO 02

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : The Ethics of the Ethics of AI

URL de élément : `https://nuage.lip6.fr/s/dBcE87XBkFnLBcx`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un article de réflexion philosophique sur l'éthique de l'IA co-écrit avec un philosophe américain et publié dans un ouvrage de synthèse important, le "Oxford Handbook of Ethics of AI". L'article est placé au tout début de cet ouvrage de synthèse ; il est d'ailleurs mentionné dans la rubrique "Ethics of Artificial Intelligence and Robotics" de la Stanford Encyclopedia of Philosophy `https://plato.stanford.edu/entries/ethics-ai/`.

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article de philosophie aborde plusieurs questions d'éthique qui devraient être préalables à toute éthique de l'IA. Ces questions se répartissent en cinq grandes catégories. La première porte sur les ambiguïtés conceptuelles des notions fondamentales de l'IA lorsqu'elles sont employées par des philosophes ou des juristes. Elles sont dues, en partie, à ce que des termes sont utilisés indifféremment en philosophie et en IA alors qu'ils ont des sens différents dans les différentes communautés. Ainsi en va-t-il par exemple, de la notion d'"agent" ou de la notion d'"autonomie".

La deuxième catégorie de questions est relative à l'estimation des risques, parfois surévalués, parfois sous-estimés. Cette question se révèle d'autant plus cruciale que l'IA-act européen se fonde sur l'anticipation de risques.

La troisième porte sur la mise en œuvre de superviseurs éthiques dans un contexte opérationnel.

La quatrième est relative à l'ambivalence de l'idée d'explication dans le cadre des systèmes d'apprentissage machine entraînés avec de très grandes masses de données. En effet, si l'IA dite explicable est très à la mode, on confond trop souvent la transparence, qui doit rendre compte fidèlement du fonctionnement d'un système, sur chaque cas particulier, au risque d'être opaque, et l'interprétation qui s'éloigne du fonctionnement effectif.

Enfin, la cinquième catégorie de question porte sur l'opposition entre deux façons de voir l'IA : une vue oppositionnelle, où l'on suppose que la machine est rivale, et une vue coopérative, où elle est un partenaire.

Le chapitre montre ensuite que beaucoup d'approches de l'éthique de l'IA n'aborde pas clairement, et de façon argumentée, ces différentes questions. Et, en conséquence, qu'il faudrait renouveler les approches de l'éthique de l'IA. Enfin, l'article aborde les difficultés à surmonter pour réaliser des superviseurs éthiques susceptibles d'imposer des prescriptions aux actions des machines.