Séminaire de la chaire éthique&IA, 1er semestre 2021-22

Les concepts de la *computer ethics*

15, 22, 29 novembre & 6 décembre 2021

Argumentaire

Nous voulons avec ce séminaire contribuer à clarifier certains concepts importants de l'éthique des algorithmes (*Computer ethics*). Cette expression, éthique des algorithmes, désigne le travail des mathématiciens-informaticiens, elle concerne la conception des algorithmes et se confond avec une double volonté exprimée par cette classe de savants et d'ingénieurs : d'une part, comprendre et maîtriser le fonctionnement leurs propres inventions, de l'autre, s'assurer de la non-dangerosité de ces dernières et fournir une garantie de leur qualité pour un usage fiable. Telle que nous l'entendons, l'expression « éthique des algorithmes » représente une des quatre branches de l'éthique de l'intelligence artificielle, avec « l'éthique artificielle », « l'éthique du numérique et de la donnée » (ou *digital ethics*), enfin « l'éthique des usages de l'IA ».

Fidèle à une tradition qui définit la philosophie comme la connaissance par concepts, le séminaire se propose d'examiner des notions très employées dans la littérature informatique consacrée à l'IA, et dont il faut clarifier le rôle. Ces concepts sont par exemple l'explicabilité, l'interprétabilité, (*Interpretability*), l'*Accountability* ou capacité d'un système d'IA à rendre des comptes, la transparence (*Transparency*), la fiabilité (*Trustworthiness*), la confiance (*Trust*), la responsabilité (dans l'expression *Responsible Artificial Intelligence*) ou encore l'équité (*Fairness*). S'ils font aujourd'hui partie intégrante des éléments techniques de la littérature académique spécialisée, nous souhaitons pour notre part braquer sur eux le regard de la philosophie, et cela pour trois raisons qui motivent ce séminaire.

La première raison se fonde sur le fait que les concepts de l'éthique algorithmique – cette discipline récente – apparaissent « jeunes » : s'ils se rodent dans leur emploi technique à partir des cas de l'algorithmique, il peut également s'avérer utile de les affuter par d'autres regards. Celui de la philosophie, « vieille » discipline accoutumée à définir et à caractériser les concepts (notamment ceux de l'éthique, qu'elle soit fondamentale ou appliquée, générale ou « régionale »), peut y contribuer.

Deuxièmement, l'apport de la philosophie se justifie également du fait que, dans leur très grande majorité, les concepts techniques de la science de l'intelligence artificielle possèdent également une acception courante qui, parce qu'elle semble intuitivement parlante, est susceptible de produire deux effets. Le premier effet revient à créer de la confusion à partir des ambiguïtés qui existent entre la signification courante et la signification technique ; le second, pour les non-spécialistes de l'IA, est de faire perdre de vue leur signification technique, laquelle conditionne pourtant l'accès au sens que leur donnent les concepteurs de l'IA. C'est le second service que peut rendre la philosophie à la science de l'IA : préciser ces deux niveaux de signification, en soulignant les aspects sur lesquels est susceptible de se produire de la confusion.

Troisièmement, le risque de confusion entre les acceptions courantes et techniques ne représente pas, toutefois, une chose négative : d'intéressantes ambiguïtés existent entre les deux niveaux de signification. Elles peuvent en effet être parlantes à la fois en ce qui concerne les manières dont les concepteurs de l'IA s'expriment à propos de leur travail scientifique, et relativement à l'horizon d'attente de la société à l'égard de l'IA. En d'autres termes, il faut voir la *computer ethics* comme un langage, comme le langage avec lequel les producteurs d'IA parlent à la société (prescripteurs, pouvoirs publics, usagers) dans une langue qui rend cette technologie « acceptable », voire désirable. En tant qu'elle reflète à la fois les justifications de la conception des algorithmes et l'horizon social d'attente à leur égard, l'étude des concepts de la *computer ethics* permet donc de comprendre de l'intérieur le projet sociétal de l'IA.