PHI-7701 Sujets spéciaux
Philosophie et éthique de l'intelligence artificielle

Professeur: Jocelyn Maclure

jocelyn.maclure@fp.ulaval.ca

I. BUT DU COURS

L'intelligence artificielle (IA) est présentement la technoscience qui suscite le plus d'intérêt à travers le monde. Après quelques décennies d'attentes déçues et de stagnation relative, ce qui a été appelé l'« hiver » de l'IA semble maintenant terminé. Des ordinateurs sont parvenus à vaincre les meilleurs joueurs humains à des jeux comme les échecs, Jeopardy et le jeu de go. Des algorithmes d'apprentissage automatique (*machine learning*) sont plus efficaces que des médecins spécialistes pour poser certains diagnostics. Des véhicules complètement autonomes pourraient rouler sur nos routes dans la prochaine décennie. Plusieurs avancent que la capacité croissante des systèmes d'IA à percevoir correctement le monde extérieur, à traiter les langues naturelles et à établir des relations significatives entre des données nombreuses et variées (*Big Data*) fera en sorte que des tâches exécutées par des humains seront de plus en plus confiées à des machines, ce qui pourrait engendrer un « chômage technologique » important, ainsi qu'une remise en question de la place du travail rémunéré dans la vie humaine.

Les progrès récents en IA ont relancé les spéculations sur l'émergence d'IA « fortes » et « générales » dotées d'une intelligence supérieure à celle des êtres humains. Certaines des meilleures fictions actuelles—*Her*, *Ex Machina*, *Black Mirror*, *Westworld*, le roman d'Ian McEwan *Machines Like Me*—explorent à nouveaux frais le thème de la relation entre l'être humain et des IA capables de réussir aisément le « jeu de l'imitation » (ou « test de Turing »). Des philosophes, chercheurs et personnalités publiques influentes comme Nick Bostrom, David Chalmers, Max Tegmark, Ray Kurzweil, Stephan Hawking, Stuart Russell, Bill Gates et Elon Musk soutiennent (avec plus ou moins de nuance selon les cas) que l'émergence de « superintelligences » artificielles est possible et que ces IA pourraient poser un « risque existentiel » eu égard à la survie même du genre humain.

Les discours inflationnistes ne manquent dans l'univers de l'IA : « Superintelligence », « singularité », « risque existentiel », création d'un « cerveau complet artificiel », émergence d'une conscience « sans substrat biologique », statut moral et droits des robots, « fin du travail », « Quatrième Révolution industrielle », « obsolescence de l'être humain ». Sans faire preuve de dogmatisme, ces perspectives seront passées au crible d'une approche analytique, critique et déflationniste dans le cadre du séminaire.

Cela étant dit, le sain scepticisme au sujet des perspectives inflationnistes portées par le battage médiatique actuel (*hype*) ne doit pas nous faire perdre de vue les

enjeux philosophiques, éthiques et politiques complexes soulevés par les avancées actuelles de l'IA. Nous aborderons des questions comme :

- Une IA pourrait-elle être consciente? Une intelligence générale et multidimensionnelle comme celle démontrée par les humains peut-elle être créée artificiellement? Le corps et l'enchâssement dans un «monde vécu» sont-ils, comme le veut la tradition phénoménologique, nécessaire à la cognition véritable? Quelles sont les forces et limites de l'approche inductive, probabiliste et corrélationniste du nouveau paradigme scientifique en IA (apprentissage automatique, réseaux de neurones artificiels)?

- Quel devrait être le statut moral et juridique des agents artificiels? Devrait-on leur accorder des droits? Doivent-ils être tenus responsables des conséquences de leurs erreurs ou défaillances (pensons à une information erronée donnée par un assistant vocal ou à un accident causé par un véhicule automne)? Les processus décisionnels des algorithmes d'apprentissage profond sont, contrairement aux approches classiques en IA basées sur la logique formelle, opaques. Doit-on exiger qu'ils puissent expliquer leurs décisions et justifier leurs jugements?

- Quelles sont les conséquences de la gouvernance algorithmique de nos vies individuelles et collectives? Des algorithmes d'aide à la décision sont de plus en plus utilisés pour déterminer qui sera reçu en entrevue d'embauche, admis dans un programme d'étude, admissible à une libération conditionnelle, etc. Les machines peuvent-elles prendre des décisions discriminatoires? Pourquoi protège-t-on la vie privée et comment l'IA la menace-t-elle? Les algorithmes utilisés par les réseaux sociaux contribuent-ils à la création de « bulles informationnelles », à la détérioration de la qualité épistémique du débat public et à la polarisation sociale? L'automatisation engendrée par l'IA dans les milieux de travail fournit-elle un argument supplémentaire aux défenseurs de l'idée d'un revenu minimum garanti inconditionnel?

II. OBJECTIFS

Objectifs de connaissance et de compréhension

a. Comprendre les théories et thèses des auteur-e-s;
b. Comprendre ce qu'est l'IA, les raisons de sa renaissance, ainsi que les forces et limites du nouveau paradigme dominant;
c. Cerner les enjeux philosophiques, éthiques et politiques soulevés par l'intégration de systèmes d'IA dans les différentes sphères de la vie humaine.

Objectifs d'habiletés intellectuelles

a. Être capable de mettre en relation et de faire dialoguer des théories contradictoires ou complémentaires;
b. Être capable de distinguer et de mettre en relation des problématiques qui relèvent de différents champs de la philosophie (philosophie de l'esprit, sciences cognitives, éthique, philosophie politique).

c. Développer la capacité de réfléchir de façon normative à tous les niveaux de généralité (de l'éthique normative fondamentale à l'éthique normative appliquée);
d. Développer sa capacité d'exprimer sa pensée de façon claire, de débattre et participer à la recherche conjointe de jugements rigoureux sur les questions abordées dans le séminaire;
e. Permettre à l'étudiant de clarifier et nuancer ses positions personnelles face aux diverses questions liées à l'IA et, ce faisant, de mieux comprendre l'esprit humain et le monde dans lequel il vit.

III. CONTENU

La liste des lectures hebdomadaires sera remise à la première séance du séminaire

IV. FORMULE PÉDAGOGIQUE

La formule retenue sera celle du *séminaire de recherche*. Le but de cette formule est de préparer l'étudiant.e à la recherche, à la rédaction et à la présentation orale des idées. L'étudiant.e aura à présenter en classe une synthèse des lectures obligatoires, ainsi qu'à réagir aux exposés de ses collègues. Le professeur fera de brèves introductions en début de séance et interviendra ponctuellement afin d'expliquer certaines thèses, préciser des faits, présenter des arguments et recadrer la discussion. L'objectif ultime est qu'étudiant.e.s et professeur enrichissent leur compréhension des théories et des phénomènes étudiés par la médiation des lectures et des discussions hebdomadaires. La lecture attentive des textes est une condition essentielle au bon déroulement du séminaire. Les étudiants devront déposer de courts commentaires sur les lectures obligatoires sur le forum ENA.

V. LECTURES OBLIGATOIRES ET LECTURES SUGGÉRÉES

Voir section III

VI. MODE ET CRITÈRES D'ÉVALUATION

1) 8 brefs commentaires critiques des lectures obligatoires (une page ou moins). Le commentaire critique doit cerner, de façon concise, le ou les principaux problèmes soulevés par la thèse de l'auteur étudié et poser quelques questions sur lesquelles nous pourrons revenir dans la période de discussion. Les commentaires doivent être déposés sur le forum le jour précédant le séminaire. 20% de la note finale.

2) Une présentation de 20-25 minutes en classe. Les présentations portent sur les lectures obligatoires et ont pour but de lancer les discussions de groupe sur les textes à l'étude.

L'étudiant-e doit présenter l'approche de l'auteur étudié, son but, expliquer (lorsque pertinent) les concepts utilisés, faire une synthèse de l'argumentaire et soulever des pistes de réflexion critique. 20% de la note finale.

3) Un plan détaillé du travail de recherche, incluant une présentation de la problématique et du but du texte, la structure provisoire et une courte bibliographie (livres, articles publiés dans des revues universitaires et chapitres de livres). 10% de la note finale. Date de remise : 29 novembre par courriel.

4) Travail de recherche de 10 pages (double interligne) sur un thème pertinent dans le cadre du séminaire. Critères de correction : compréhension de la problématique, qualité de la recherche, clarté et rigueur de l'argumentation et qualité de la langue. 40% de la note finale. Date de remise : 20 décembre par courriel.

5) Participation en classe. 10% de la note finale.

POLITIQUES GÉNÉRALES

Notation selon l'échelle en vigueur à la Faculté de philosophie.

Des points seront enlevés pour les incorrections de la langue (voir *Politique du français* disponible sur le site Web de la Faculté de philosophie).

Le plagiat est tout à fait proscrit. Se référer au site Internet de la Faculté et au *Règlement des études*.

Étudiants ayant un handicap : Voir la *Procédure de mise en application des mesures d'accommodations scolaires*, à l'adresse suivante :

http://www.aide.ulaval.ca/cms/site/cocp/pid/1936