PHI 6385 A22 Esprits Numériques

Prof. :	Jonathan Simon
Bureau :	2910 Édouard-Montpetit, 4e étage, local 428
Téléphone :	514-343-6111 #42997
Communication :	sur StudiUM
Cours :	Mardi 13h30 – 16h30
Mode d'enseignement :	presentiel
Examen Final :	Pas d'examen final
Heure de disponibilité :	sur rendez-vous

I. Objectifs

Ce séminaire sera une plongée profonde dans les questions entourant la conscience artificielle. Les IA du futur seront-elles conscientes ? Qu'en est-il de celles d'aujourd'hui ? Comment pouvons-nous le savoir ? En quoi pourrait la conscience artificelle être semblable, ou différente, de la conscience humaine, et que pouvons-nous apprendre sur la conscience en général en y réfléchissant ?

Le séminaire abordera des textes de la philosophie de l'esprit et des sciences cognitives, ainsi que des textes de neuroscience et d'intelligence artificielle. Le cours présuppose une familiarité avec les méthodes et les pratiques de la philosophie analytique, mais aucune formation scientifique n'est requise.

Le séminaire sera divisé en deux parties. La première partie abordera des questions générales de méthodologie en philosophie de l'esprit, et de la possibilité de la conscience des machines. Dans la deuxième partie, nous explorerons des questions plus spécifiques sur les formes que pourrait prendre une telle conscience mécanique, en accordant une attention particulière aux capacités des réseaux neuronaux profonds contemporains.

II. Modes d'évaluation

L'évaluation des étudiants comporte les composantes suivantes:

- Une présentation en classe (avec diapositives) d'environ 40 minutes, et la conduite de la discussion qui s'ensuit (45 minutes)
  25%
- Une dissertation de 1500-2500 mots
  25%
- Une dissertation finale de 1500 – 6000 mots
  50%

1ère dissertation:	**à remettre le 30 octobre 2022 (25%)**
2ème dissertation:	**à remettre le 16 décembre 2022 (50%)**

- Pénalités : 5% par jour ouvrable de retard; 5% par 100 mots *au-delà* de la limite de 2500/6000 mots (première/deuxième diss.). *Note : j'accorde de l'importance à l'économie et à l'efficacité dans la rédaction, et par conséquent je ne pénalise pas les travaux trop courts. Cependant, gardez à l'esprit qu'il est plus difficile d'atteindre tous les objectifs du devoir si vous êtes trop économe. Je ne recommande pas de soumettre un travail de moins de 1500 mots, à moins que vous ne sachiez vraiment ce que vous faites.*
- Politique de contestation : pendant une semaine aprèes le remise du travail vous pouvez demander une réévaluation, mais notez que la note peut changer dans les deux sens quand une réévaluation est demandée.

III. Document de référence

- Toutes les lectures seront disponibles sur le site web du cours (https://jonsimon.net/esprits-numeriques).
  [les lectures sont protégées par un mot de passe : contactez-moi]

Une bibliographie détaillée ainsi que le calendrier des lectures sont inclus ci-dessous.

IV. Plagiat

Le plagiat à l'U de M est sanctionné par le Règlement disciplinaire sur la fraude et le plagiat concernant les étudiants. Pour plus de renseignements, consultez le site www.integrite.umontreal.ca.
Pour des clarifications sur la définition du plagiat, je vous recommande également le site http://web.mit.edu/academicintegrity/

Nous vous invitons à consulter le document qui formule les **lignes directrices sur le climat du département et la lutte contre le harcèlement :**
https://philo.umontreal.ca/departement/comite-acces-a-legalite-et-climat/

V. La propriété intellectuelle et le droit à l'image

L'usage de tout document déposé sur StudiUM pour chaque cours (incluant les enregistrements audio et vidéo) est assujetti à l'engagement de chaque étudiant à respecter la propriété intellectuelle et le droit à l'image. Il est interdit de faire une captation audio ou vidéo du cours, en tout ou en partie, sans le consentement écrit du professeur. Le non-respect de cette règle peut mener à des sanctions disciplinaires en vertu de l'Article 3 du Règlement disciplinaire concernant les étudiants.

VI. Plan détaillé du cours

I. INTRO : CONSCIENCE VS INTELLIGENCE

6 Septembre: La conscience des invertébrés

- Birch, J. « The Search for Invertebrate Consciousness »
- Barron, A., et Klein, C. « Insects have the capacity for subjective experience »

II. LA CONSCIENCE EST-ELLE COMPUTATIONELLE?

13 Septembre: Qu'est-ce que le calcul (computation) ? Qu'est-ce que c'est que de mettre en œuvre une fonction (de calcul) ?

- Turing, A. « Computing Machinery and Intelligence »
- Maudlin, T.. « Computation and Consciousness »

Recommandé:

- Marr, D.. « Vision » *extraits*
- Klein, C. « Olympia and other O-Machines »
- Chalmers, D. « Does a Rock Implement every Finite-State Automaton? »
- Bostrom, N. « Brain Duplication and Degrees of Consciousness »

20 Septembre: Arguments pour la possibilité d'une conscience de machine

- Chalmers, D. « Organizational Invariance », *extraits*
- Putnam, H. « The Nature of Mental States »

27 Septembre: Arguments contre la possibilité d'une conscience de machine

- Block, N. « Troubles with Functionalism »
- Searle, J. « Minds, Brains and Programs »
- Godfrey-Smith, P. « Minds, Machines and Metabolism »
- Seth, A. « Being You » *extraits*

4 Octobre: Quels algorithmes, partie 1 : Connectionisme vs Systèmes de Symboles

- Fodor J. et Pylyshyn, Z. « Connectionisme et architecture cognitive: une analyse critique »
- Smolensky, P. « Le traitement approprié du connectionisme »

Recommandé:

- Dreyfus, H. « What Computers Still Can't Do » *extraits*

---

11 Octobre: Comment savoir ? Sur les tests et la méthodologie de la découverte

- Schneider, S. « Artificial You » *extraits*
- Udell, B. et Schwitzgebel, E. « Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed »

Recommandé :

- Shevlin, H. « How Could We Know When a Robot was a Moral Patient? »
- Shanahan, M. « Beyond Humans, What Other Kinds of Minds Might be Out There? »
- Birch (de la première semaine)

---

18 Octobre: Théories de la conscience : Un aperçu, et plus sur la méthodologie

- Seth, A. et Bayne, T. « Theories of Consciousness »
- Dehaene, S.,Lau, H., et Kouider, S.. « What is Consciousness, and Could Machines Have It? »

Recommandé

- Shevlin, H. « Non-Human Consciousness and the Specificity Problem »
- Bayne, T. et Shea, N. « Consciousness, Concepts and Natural Kinds »
- Birch, J., Ginsburg, S., et Jablonka, E. « Unlimited Associative Learning and the Origins of Consciousness »

---

25 Octobre: Semaine de Lecture

III. LES CAPACITÉS (ET LES LIMITES) ÉTONNANTES DES RÉSAUX NEURONAUX PROFONDS

1 Novembre: Une introduction technique aux réseaux neuronaux, de l'approximation universelle à la back propagation

- Nielsen, M. « Neural Networks and Deep Learning, Ch. 1, Ch. 4, Ch. 5 »

Recommandé:

- Nielsen, Ch. 6
- Goodfellow, I., Bengio, Y. et Courville, A. « Deep Learning »

---

8 Novembre: Vision, perception et réseaux neuronaux convolutionnels (CNN); Pensée, sens et réseaux neuronaux récurrents (RNN)

- Buckner, C. « Philosophical Issues in Deep Learning »

Recommandé:

- Buckner, C. « Understanding Adversarial Examples Requires a Theory of Artifacts for Deep Learning ».
- Alammar, J. « The Illustrated Word2Vec ». (*blog)* https://jalammar.github.io/illustrated-word2vec/
- Günther, F., Rinaldi, L., et Marelli, M. « Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions »
- Landuaer, T. et Dumais, S. « A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge »

---

15 Novembre: Désir, agence et apprentissage par renforcement

- Haas, J. « Philosophical Issues in Reinforcement Learning »

Recommandé:

- Berridge, K. et Kringelback, M. « Affective Neuroscience of Pleasure: Reward in humans and animals »
- Schroeder, T. et Arpaly, N. « The Reward Theory of Desire in Moral Psychology »
- Bickle, J. « Motivation, Decision-making and Neuorethics »
- Silver, D. et. al. « Reward is Enough »
- Amodei, D. et. al. « Concrete Problems in AI Safety »

22 Novembre : Penser, faire attention et prédire (Transformers et modèles génératifs)

Recommandé

- Vaswani, et. al. « Attention is all you need »
- Bommasani et. al. « On the Opportunities and Risks of Foundational Models »
- Elhage, et. al. « Softmax Activation Units »
- Google, « Background: What is a Generative Model?»

---

29 Novembre: D'ici à la conscience : Mise à l'échelle, architecture, données - ou un changement complet de direction ?

- Branwen, G. « The Scaling Hypothesis » (*blog)*
- Tay, Y. et. al. « Scaling Laws vs Model Architectures: How Does Inductive Bias Influence Scaling? »

Recommandé

- Hoffman, J. et. al. « Training Compute-Optimal Large Language Models »
- LeCun, Y. « A Path Toward Autonomous Machine Intelligence »
- Bengio, Y. « The Consciousness Prior »
- Marcus, G. « Deep Learning is Hitting a Wall » (*vulgarisation)*
- LeCun, Y. et Browning, J. « What AI Can Tell Us About Intelligence » (*vulgarisation)*
- Marcus, G. « Deep Learning alone isn't Getting us to Human-Level AI » (*vulgarisation)*
- Marcus, G et Davis, E. « OpenAI's Language Generator has No Idea What Its Talking About » (*vulgarisation*)
- Alexander, S. « AI Size Solves Flubs » (*blog*)
- Marcus, G. . « What Does it mean when an AI fails? A Reply.» (*blog*)
- Alexander, S. « Somewhat Contra Marcus on AI Scaling » (*blog*)
- Marcus, G. « Does AI Really Need a Paradigm Shift? » (*blog*)
- Lemoine, B. « Is LAMDA Sentient? » (*blog*)

---

6 Décembre : Êtres moraux numériques : Quelques questions clés

- Schulman, C. et Bostrom, N.. « Sharing the World with Digital Minds »
- Schwitzgebel, E. et Garza, M. , « A Defense of the Rights of Artificial Intelligences »

Recommandé

- Bayne, T. « Value of Consciousness »
- Lee, A.Y. « Speciesism and Sentientism »

Sources électroniques utiles:
L'Encyclopédie Philosophique: http://encyclo-philo.fr/
The Routledge Encyclopedia of Philosophy: http://www.rep.routledge.com
The Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/