Syllabus pour le PHI 6385 A22 – Philosophie de l'esprit
Esprits numériques

Jonathan Simon
Professeur adjoint
Département de philosophie
Bureau 428, 2910 Édouard-Montpetit,
514-343-6111 #42997
jonathan.simon@umontreal.ca

Objectifs du cours

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.

Dans la première partie, nous explorerons la question de savoir comment la nature spéciale et première personne de la conscience contraint la méthodologie de son étude scientifique. Nous passerons également en revue les principales théories de la conscience, notamment les théories centrées sur la fonction de gestion des informations (comme la théorie de l'espace de travail global et la théorie de l'information intégrée), les théories centrées sur le rôle des modèles internes du soi ou du monde (comme les théories métacognitives, les théories de l'inférence active et la théorie du schéma d'attention), les théories centrées sur le rôle de l'apprentissage et les théories centrées sur l'intégration de la perception et de l'action. Nous examinerons ensuite les arguments de haut niveau pour et contre le computationalisme.

Nous commencerons la deuxième partie par un examen du débat entre les approches du connexionnisme et des systèmes experts. Nous nous plongerons ensuite dans les approches contemporaines basées sur les réseaux neuronaux (profonds), en cherchant à comprendre pourquoi ces réseaux sont si efficaces, comment ils apprennent, et si leurs succès soutiennent les néo-empiricistes ou les néo-rationalistes. Ensuite, nous explorerons les liens entre des architectures de réseaux neuronaux spécifiques et des aspects particuliers de la cognition liés à la conscience, comme la perception (réseaux convolutionnels et transformateurs), le langage naturel (réseaux neuronaux récurrents et transformateurs), le désir, la motivation et la récompense (apprentissage par renforcement) et l'attention (mécanismes attentionnels). Enfin, nous nous

appuierons sur ce qui précède pour évaluer la probabilité que les IA du futur proche soient conscientes, et nous discuterons des implications éthiques.

Organisation

Le cours sera un séminaire. Si la taille de la classe n'est pas trop grande, la structure de chaque séance sera la suivante : pendant la première moitié de la séance, un étudiant présente / mène la conversation. Ensuite une pause, puis pendant la dernière heure, je anime la conversation.

Bibliographie indicative

Pour l'ensemble du cours, les ouvrages suivantes peuvent être consultés :

- Bayne, Tim & Shea, Nicholas (2020). Consciousness, Concepts and Natural Kinds. *Philosophical Topics* 48 (1):65-83.
- Bengio, Yoshua (2019). The Consciousness Prior.
- Birch, Jonathan (2022). The Search for Invertebrate Consciousness. *Noûs* 56 (1):133-153.
- Birch, Jonathan ; Ginsburg, Simona & Jablonka, Eva (2021). The Learning-Consciousness Connection. *Biology and Philosophy* 36 (5):1-14.
- Block, N. « Consciousness, Accessibility, and the mesh between Psychology and Neuroscience »
- Buckner, Cameron (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese* (12):1-3
- Buckner, C. et Garson, J. « **connectionism** », dans Stanford Encylopedia of Philosophy
- Chalmers, D. 2002. Philosophy of Mind: Classical and Contemporary Readings. NY: Oxford University Press.
- Crump, Andrew & Birch, Jonathan (2021). Separating Conscious and Unconscious Perception in Animals. *Learning and Behavior* 49 (4).
- Dreyfus, H. « What Computers Still Can't Do »
- Fisette, D. et P. Poirier (dir. publ.) 2002/3. Textes clés en Philosophie de l'esprit. Paris: Vrin.
- Fodor, J. et Pylyshyn, Z. « Connexionnisme et architecture cognitive : Une analyse critique »
- Goodfellow, I.; Bengio, Y. et Courville, A., «Deep Learning», MIT Press.
- Haas, J, « Reinforcement Learning: A Guide for Philosophers of Mind», Philosophy Compass
- Haugeland, J. 1985. L'esprit dans la machine. trad. franç. J. Henry. Paris: O. Jacob, 1989.
- Huxley, T.H. « On the Hypothesis that Animals are Automata, and its History »
- Klein, C. et Barron, A. « Insects have the Capacity for Subjective Experience »
- Kouider, de Gardelle, Dupoux. «Partial Awareness and the Illusion of Phenomenal Consciousness"»
- Lake, B. et Baroni, M. « Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks
- Lau, H et Persaud, N. «Broken Telephone in the Brain: the need for Metacognitive Measures»
- Naccache,L et Dehaene, S. «Reportability and Illusions of Phenomenality in the Light of the Global Neuronal Workspace Model»
- Nielsen, M. «Neural Networks and Deep Learning»
- Prinz, J. «Accessed, Accessible, Inaccessible: Where to draw the Phenomenal line»
- Searle, J. « Can Computers Think? »
- Seth, A. K. (2009). Explanatory correlates of consciousness: Theoretical and computational challenges. Cognitive Computation, 1 (1), 5063.
- Seth, A. et Bayne, T. «Theories of Consciousness» ,Nat Rev Neurosci, 2022
- Smolensky, P. « Le traitement approprié du connexionnisme »
- Turing, A. « Les ordinateurs et l'intelligence »
- Tye, M. « The Problem of Simple Minds: Is there Anything it is Like to be a Honeybee? »