Département de philosophie

SYLLABUS PHI 3330

Théories philosophiques de l'IA

Objectifs généraux du cours :

L'objectif de ce cours est d'explorer les questions philosophiques fondamentales liées à la théorie de l'intelligence artificielle, en mettant particulièrement l'accent sur les intersections avec les sciences cognitives. Nous chercherons à répondre à des questions clés : pourquoi les systèmes d'intelligence artificielle actuels sont-ils aussi performants ? Quelles sont les limites qui expliquent qu'ils ne le soient pas davantage ? Comment l'étude de l'IA peut-elle éclairer notre compréhension de l'esprit et du cerveau ? Et réciproquement, comment les connaissances issues des sciences cognitives peuvent-elles enrichir notre conception et nos théories de l'IA ?

Le cours couvrira un large éventail de sujets : du débat classique entre les approches symboliques et connexionnistes à l'évolution de ces dernières, des perceptrons aux modèles d'apprentissage profond. Nous aborderons également des questions contemporaines comme le débat sur le *scaling*, les algorithmes de représentation vectorielle, les mécanismes d'attention, la convolution, la récurrence, et l'apprentissage par renforcement. D'autres thématiques incluront la créativité et l'agentivité des systèmes d'IA, les enjeux liés à leur alignement et leur interprétabilité, ainsi que les réflexions sur la nature de l'intelligence elle-même.

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? »