PHI 3330 H25 Théories philosophiques de l'IA

Prof. :	Jonathan Simon
Bureau :	2910 Édouard-Montpetit, 3e étage, local 322
Communication :	jonathan.simon@umontreal.ca
Cours :	mardi 12h30 – 15h29
Mode d'enseignement :	En présentiel
Local :	B-440 Pav Marie Victorin (avant semaine de relâche) B-4340 Pav Jean Brillant (après semaine de relâche)
Examen Final :	Pas d'examen final
Heure de disponibilité :	sur rendez-vous

I. Objectifs

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 pourquoi 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.

II. Modes d'évaluation

La formule pédagogique du cours est le cours magistral.

L'évaluation des étudiants comporte les composantes suivantes:

- Deux dissertations de 1000-1500 mots, **à remettre sur StudiUM**. (des versions papier ne seront pas acceptées).
- Pour chaque dissertation, un formulaire (**à remettre sur StudiUM**) décrivant votre plan de rédaction.

Formulaire pour 1ère dissertation:	**à remettre le 24 fév 2025 (2%)**
1ère dissertation:	**à remettre le 3 mars 2025 (38%)**
Formulaire pour 2ème dissertation :	**à remettre le 17 avril 2025 (2%)**
2ème dissertation:	**à remettre le 24 avril 2025 (58%)**

- Pénalités : 5% par jour ouvrable de retard; 5% par 100 mots au-delà ou en-dessous de la limite de 1500-2000 mots.
- Politique de contestation : 1) discussion avec votre correcteur 2) demande de réévaluation auprès du professeur – notez que les deux étapes ne sont possibles que pendant une semaine après la remise du travail et que la note peut changer dans les deux sens quand une réévaluation est demandée.

III. Document de référence

- **Littérature principale (obligatoire) :**

- Lectures disponibles sur le site web du cours (https://jonsimon.net/theories_IA) [certains des 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 https://integrite.umontreal.ca/accueil/

Pour des clarifications sur la définition du plagiat, je vous recommande également le site http://web.mit.edu/academicintegrity/

Pour obtenir des informations sur la politique de l'université en matière de violence, de harcèlement et de discrimination, consulte le site web du Bureau du respect de la personne **: https://respect.umontreal.ca/accueil/**

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

Partie I : QUESTIONS PRÉLIMINAIRES

---

14 janvier: Qu'est-ce qu'un ordinateur ? Qu'est-ce qu'un algorithme ?
Littérature principale:

- Turing, A. « Machines informatiques et intelligence » [français], [anglais]

Littérature recommandé:

- Chalmers, D. « Do Dust Clouds Run Computer Programs » Reality+, Ch. 21 (sur STUDIUM)
- Marr, D. « Vision » ch. 1. [anglais]

---

21 janvier: Quels algorithmes? Connectionisme vs Systèmes de Symboles
Littérature principale:

- Fodor, J. et Pylyshyn, Z. « Connexionnisme et architecture cognitive : Une analyse critique » [français], [anglais]

Littérature recommandé:

- Smolensky, P. « Le traitement approprié du connexionnisme » [français]
- Jake Quilty-Dunn, Nicolas Porot et Eric Mandelbaum (2022) « The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences » [anglais]

Partie II : LES CAPACITÉS (ET LES LIMITES) ÉTONNANTES DES RÉSEAUX NEURONAUX PROFONDS

---

28 janvier: Une légère introduction technique aux réseaux neuronaux, de l'approximation universelle à la back propagation (représentation et apprentissage)
Littérature principale:

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

Littérature recommandé:

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

---

4 février: Division du travail, Invariance de la translation et perception dans les réseaux neuronaux convolutionnels (CNN)
Littérature principale

- Buckner, C. « Deep Learning: A Philosophical Introduction » [anglais]

Littérature recommandé:

- Buckner, C. « Understanding Adversarial Examples Requires a Theory of Artifacts for Deep Learning », [anglais]
- Yann Lecun, Leon Bottou, Yoshua Bengio et Patrick Haffner « Gradient-Based Learning Applied to Document Recognition » [anglais]

---

11 février: La compréhension en tant que similarité : la représentation vectorielle sémantique

Littérature principale:

- Piantadosi, S. et Hill, F. « Meaning without reference in large language models », [anglais]
- David Chalmers (2023) « Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models »,[anglais]

Littérature recommandé:

- Alammar, J. « The Illustrated Word2Vec » (blog)
- Prinz, J. « Empiricism and State Space Semantics » [anglais]
- Günther, F., Rinaldi, L., et Marelli, M. « Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions », [anglais]
- Daniel Jurafsky & James H. Martin. (2024) « Speech and Language Processing, Ch. 6. » [anglais]
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent et Christian Jauvin. (2003) « A Neural Probabilistic Language Model » [anglais]

---

18 février : Penser, faire attention et génerer (Transformers et modèles génératifs)

Littérature principale:

- Lindsay, G. « Attention in Psychology, Neuroscience and Machine Learning », [anglais]

Littérature recommandé:

- Vaswani, et. al. « Attention is all you need », [anglais]
- Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio (2014), « Neural Machine Translation by Jointly Learning to Align and Translate » [anglais]
- Jiang, L. « How GPT Works: A Metaphoric Explanation of Key-Value-Query in Attention Using a Tale of Potion», [anglais]
- Google, « Background: What is a Generative Model?», [anglais]
- Søgaard, A. « Understanding Models Understanding Language »,[anglais]
- Landgrebe, J. et Smith, B. « Why Machines do not Understand: A Reply to Søgaard »,[anglais]
- Elhage et. al. (2021) « A Mathematical Framework for Transformer Circuits »,[anglais]

---

25 février: Agentivité, créativité et apprentissage par renforcement

Littérature principale:

- Julia Hass « Reinforcement Learning for Philosophers » [anglais]
- Lindsay Brainard, « The curious case of uncurious creation » [anglais]

Littérature recommandé:

- Berridge, K. et Kringelback, M. « Affective Neuroscience of Pleasure: Reward in humans and animals », [anglais]
- Schroeder, T. et Arpaly, N. « The Reward Theory of Desire in Moral Psychology », [anglais]
- Alex Turner (2023) « Reward is not the Optimization Target », [anglais]
- Alex Turner (2023) « Think Carefully Before Calling RL Agents Policies », [anglais]
- Silver, D. et. al. « Reward is Enough », [anglais]
- Butlin, P. « Reinforcement Learning and Artificial Agency » [anglais]

- Margaret Boden (2014) « Creativity and Artificial Intelligence : A Contradiction in Terms? », [anglais]
- Margaret Boden (2004) « Creativity and Artificial Intelligence », [anglais]
- « Eight Scholars on Art and Artificial Intelligence », [anglais]
- Nick Cave, « This Song Sucks » [anglais]
- Jürgen Schmidhuber, « Formal Theory of Fun and Creativity Explains Science, Art, Music, Humor » [anglais]

Partie III : LA NATURE DE L'INTELLIGENCE

---

11 mars: Le romantisme contre les Lumières (et le débat sur la cognition incarnée)

Littérature principale

- Daniel Dennett, « Aching Voids and Making Voids » [anglais]
- Shane Legg and Marcus Hutter, « Universal Intelligence : A Definition of Machine Intelligence » [anglais]

Littérature recommandé:

- Godfrey-Smith, P. « Minds, Machines and Metabolism », [anglais]
- Jobst Landgrebe et Barry Smith, « An Argument for the Impossibility of Machine Intelligence » [anglais]
- Shane Legg and Marcus Hutter, « A Collection of Definitions of Intelligence" [anglais]

---

18 mars: L'innéisme contre l'empirisme (et le débat sur « scaling »)

Littérature principale:

- Sutton, R. « The Bitter Lesson » [anglais]
- Gary Marcus (2018) « Deep Learning : A Critical Appraisal » [anglais]
- François Chollet (2017), « On the Measure of Intelligence » [anglais]

Littérature recommandé:

- Bubeck, S. et. al. « Sparks of Artificial General Intelligence: Early experiments with GPT-4 » [anglais]
- Bommasani et. al. « On the Opportunities and Risks of Foundational Models », [anglais]
- Branwen, G. « The Scaling Hypothesis » (blog) [anglais]
- Kaplan, J. et. al. « Scaling Laws for Neural Language Models », [anglais]
- Tay, Y. et. al. « Scaling Laws vs Model Architectures: How Does Inductive Bias Influence Scaling? », [anglais]
- Hoffman, J. et. al. « Training Compute-Optimal Large Language Models », [anglais]
- LeCun, Y. « A Path Toward Autonomous Machine Intelligence », [anglais]
- Bengio, Y. « The Consciousness Prior », [anglais]
- David H. Wolpert (2020) « What is important about the No Free Lunch theorems? » [anglais]
- Gerhard Schurz (2022) « No Free Lunch Theorem, Inductive Skepticism, and the Optimality of Meta-induction » [anglais]
- Marcus, G. « Deep Learning is Hitting a Wall » (vulgarisation), [anglais]
- LeCun, Y. et Browning, J. « What AI Can Tell Us About Intelligence » (vulgarisation), [anglais]
- Marcus, G. « Deep Learning alone isn't Getting us to Human-Level AI » (vulgarisation), [anglais]
- Marcus, G et Davis, E. « OpenAI's Language Generator has No Idea What Its Talking About » (vulgarisation), [anglais]
- Alexander, S. « AI Size Solves Flubs » (blog), [anglais]

- Marcus, G. . « What Does it mean when an AI fails? A Reply.» (blog), [anglais]
- Alexander, S. « Somewhat Contra Marcus on AI Scaling » (blog), [anglais]
- Marcus, G. « Does AI Really Need a Paradigm Shift? » (blog), [anglais]
- Lovely, G. (2024) « Is Deep Learning Really Hitting a Wall? » (blog), [anglais]
- François Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers (Dec 4 2024) « ARC Prize 2024: Technical Report » [anglais]
- Francois Chollet (Dec 20 2024) « OpenAI o3 Breakthrough High Score on ARC-AGI-Pub » [anglais]

Partie IV : AGENTS, PATIENTS, ET MONSTRES (SENTIENCE, INTERPRÉTABILITÉ, ALIGNEMENT, RESPONSIBILITÉ)

---

25 mars: Les esprits numeriques — La question de la conscience / la sentience

Littérature principale:

- Block, N. « Troubles with Functionalism » [français], [anglais]
- Dehaene, S.,Lau, H., et Kouider, S.. « What is consciousness, and could machines have it? » [anglais]
- Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, Rufin VanRullen
- Birch, J. « The Search for Invertebrate Consciousness », [anglais]

Littérature recommandé:

- Lemoine, B. « Is LAMDA Sentient? » (blog), [anglais]
- Michael H. Herzog, Michael Esfeld et Wulfram Gerstner « Consciousness & the small network argument » [anglais]
- Shanahan, M. « Beyond Humans, What Other Kinds of Minds Might be Out There? », [anglais]
- Nagel, T. « Quel effet cela fait-il d'être une chauve-souris? » [français], [anglais]
- David Chalmers (2023) « Could A Large Language Model Be Conscious? »,[anglais]
- Blaise Agüera y Arcas (2023) « Do Large Language Models Understand Us? »,[anglais]
- Campero, A. « Report on Candidate Computational Indicators for Conscious Valenced Experience », [anglais]

---

1 avril: L'interpretabilité et l'alignement

L'interpretabilité

- Beisbart, C. et Räz, T. « Philosophy of science at sea: Clarifying the interpretability of machine learning ».[anglais]
- Brent Mittelstadt (2022) « Interpretability and Transparency in Artificial Intelligence » [anglais]
- Templeton, et. al, « Scaling Monosemanticity : Extracting Interpretable Features from Claude 3 Sonnet » [anglais], [blog]
- Elhage, et. al. « Softmax Linear Units », [anglais]

L'alignement

- Bengio, Y. « How Rogue AIs May Arise ». [anglais]
- Karina Vold, Daniel R. Harris « How Does Artificial Intelligence Pose an Existential Risk? » [anglais]
- Amodei, D. et. al. « Concrete Problems in AI Safety », [anglais]

- Bengio, Y. et. al. (2024) « Managing Extreme AI Risks Amid Rapid Progress », [anglais]
- Michael K. Cohen , Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell (2024) « Regulating advanced artificial agents », [anglais]
- Blaise Agüera y Arcas, Blake Richards, Dhanya Sridhar and Guillaume Lajoie (2023) « Fears About AI Existential Risk are Overdone », [anglais]

---

8 avril: L'agentivité morale et la patientivité morale des IA

Long and Sebo report
Sharing the World, excerpts from Deep Utopia?
Alexander and Simon

- Schulman, C. et Bostrom, N.. « Sharing the World with Digital Minds », [anglais]

Powerpoint de Prof. Simon pour séance 12

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/