COMMENT SAVOIR SI LES ROBOTS SONT CONSCIENTS ?

La question de trouver une méthodologie appropriée pour la découverte

Jonathan Simon

PHI-6385

Séance 7 2024

APERCU

APERCU

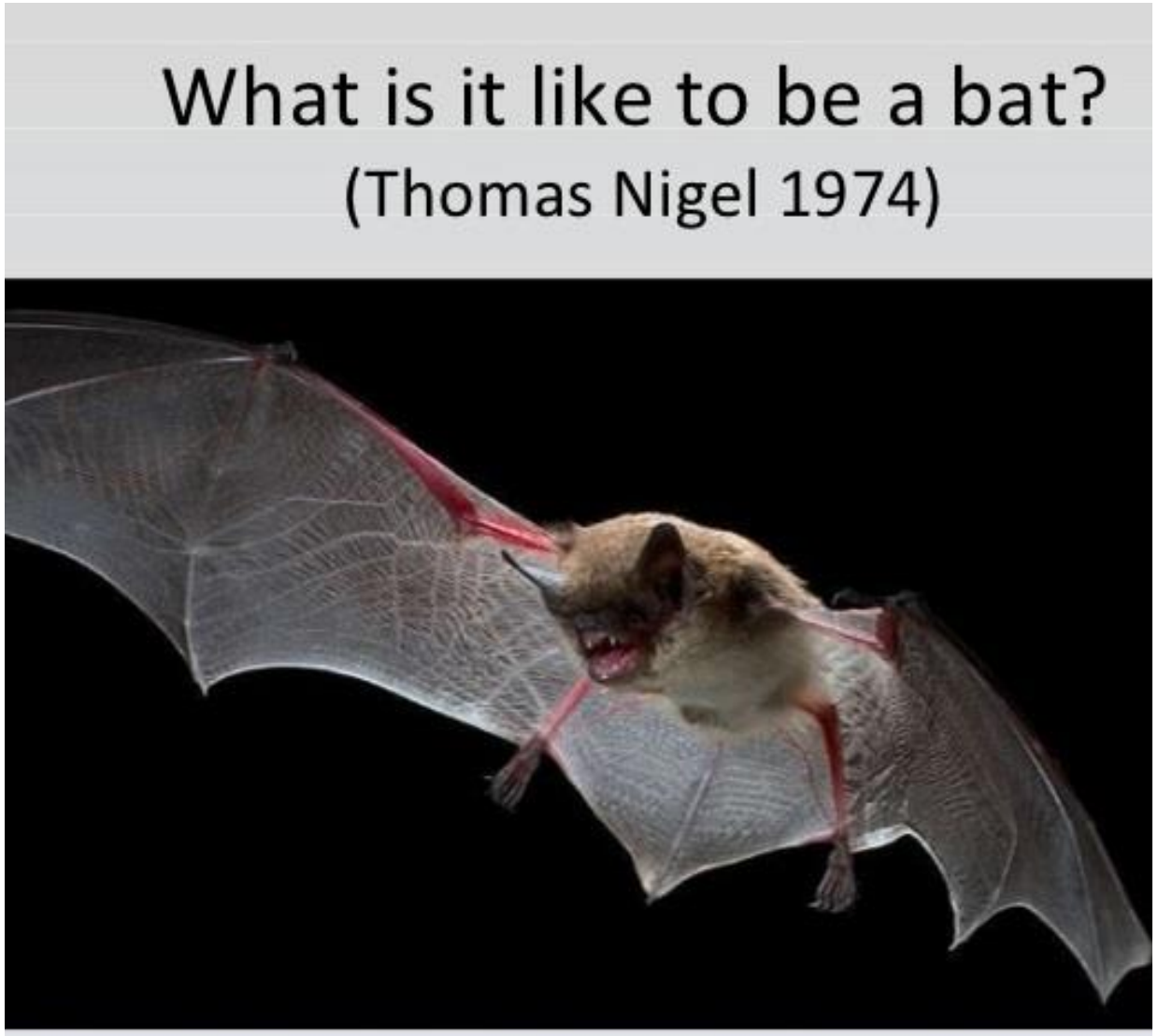
- 1) La question de la méthodologie
- 2) Méthodologie neutre sur le plan théorique «Theory-neutral methodology» (Schneider, Shevlin 2021a, Tye)
- 3) Méthodologie lourde sur le plan théorique «Theory-heavy methodology» (Butlin et. al.?)
- 4) Méthodologie legère sur le plan théorique «Theory-light methodology»
- (Birch, Andrews, Dung, Shevlin 2021b, Bayne et Shea, Butlin et. al.?)

LA QUESTION DE LA METHODOLOGIE

METHODOLOGIE

- Nous avons examiné les arguments généraux pour et contre la possibilité de la conscience de la machine et nous avons examiné deux types très généraux de conception de système qui pourraient soutenir la conscience de la machine.
- ... mais le mystère reste entier quant aux objets qui pourraient en être dotés (et à ce qu'ils pourraient ressentir).

METHODOLOGIE



METHODOLOGIE

- L'article de Nagel est une excellente illustration du type de question qui reste ouverte

METHODOLOGIE

- Dans son article « La recherche de la conscience des invertébrés », Jonathan Birch distingue trois approches méthodologiques pour déterminer si des systèmes qui ne nous ressemblent pas sont conscients :

METHODOLOGIE

- 1) Méthodologie neutre sur le plan théorique «Theory-neutral methodology»
- 2) Méthodologie lourde sur le plan théorique «Theory-heavy methodology»
- 3) Méthodologie legère sur le plan théorique «Theory-light methodology»

Photo Credit: CinemaBlend

METHODOLOGIE

- Théorie dans quel sens ?
- Nous entendons par là : les théories du principe de corrélation / d'identité psychophysique.
- (Théorie de l'espace de travail global, théorie de la pensée d'ordre supérieur, etc …).

METHODOLOGIE

- 1) Méthodologie neutre sur le plan théorique «Theory-neutral methodology»

- Appel à des méthodes que nous pouvons accepter indépendamment de toute hypothèse théorique (tests comportementaux ? Auto-évaluation ?)

METHODOLOGIE

- 2) Méthodologie lourde sur le plan théorique «Theory-heavy methodology»
- Détermine d'abord quelle théorie est vraie (ou, fixe une distribution de probabilité), puis utilise-la pour déterminer (une distribution de probabilité pour) si un système d'IA donné est conscient, selon la théorie en question.

METHODOLOGIE

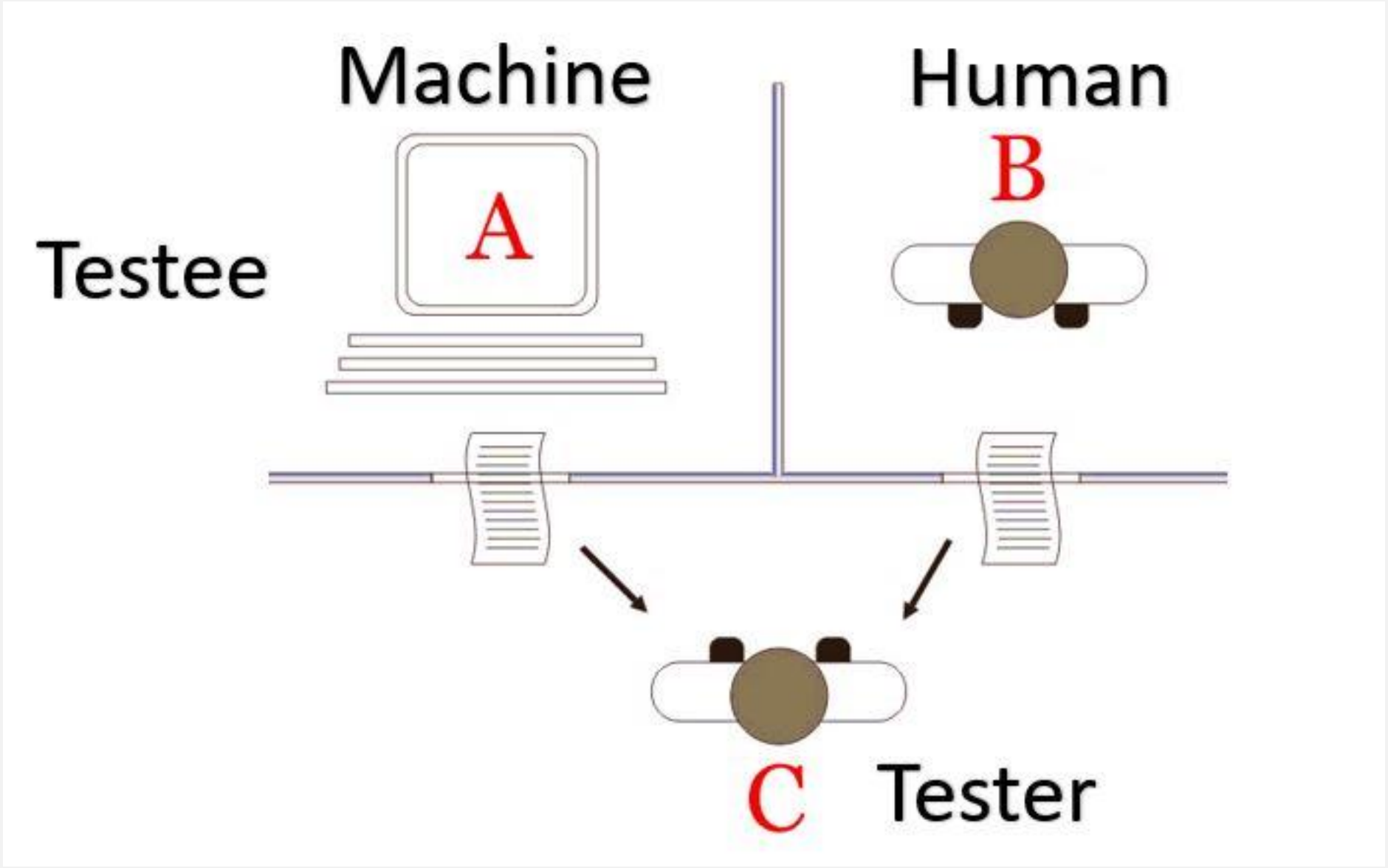
3) Méthodologie legère sur le plan théorique «Theory-light methodology»

Un compromis entre 1 et 2 : utiliser la théorie pour identifier les marqueurs, puis rechercher les grappes / types naturels qui accompagnent ces marqueurs.

THEORY NEUTRAL

THEORY NEUTRAL

- *Locus Classicus: Turing – le test de Turing.*
- *Strictement, son test était un test d'intelligence (et beaucoup de choses que nous considérons comme conscientes, comme les chiens ou les nourrissons, n'y parviennent pas)...*


Machine
Human
B
A
Testee
C
Tester

THEORY NEUTRAL

- *Mais nous pouvons le considérer comme un test de suffisance pour la conscience*
- *Le problème : ce n'est pas neutre, cela entre en conflit avec de nombreuses théories de la conscience (qui disent que la conscience dépend d'une architecture qu'un chatbot passant le test de Turing pourrait ne pas avoir)*

THEORY NEUTRAL

- *Schneider:*

- *Deux nouveaux tests*
- *1) Le test de conscience de l'IA (ACT)*
- *2) Le test de la puce*

THEORY NEUTRAL

- *I) Le test de conscience de l'IA (ACT)*

- *Le système développe-t-il de manière autonome la capacité ou la tendance à parler de ses états internes, même si nous ne l'avons pas entraîné à le faire ?*

THEORY NEUTRAL

- *2) Le test de la puce*

- *(moins praticable) : remplacez vos propres neurones par des puces de silicium et voyez comment vous vous sentez.*

THEORY NEUTRAL

- *Shevlin:*

- *Test d'équivalence cognitive :*
- *Évaluer si un système possède les mêmes capacités cognitives que celles que nous associons à la conscience.*

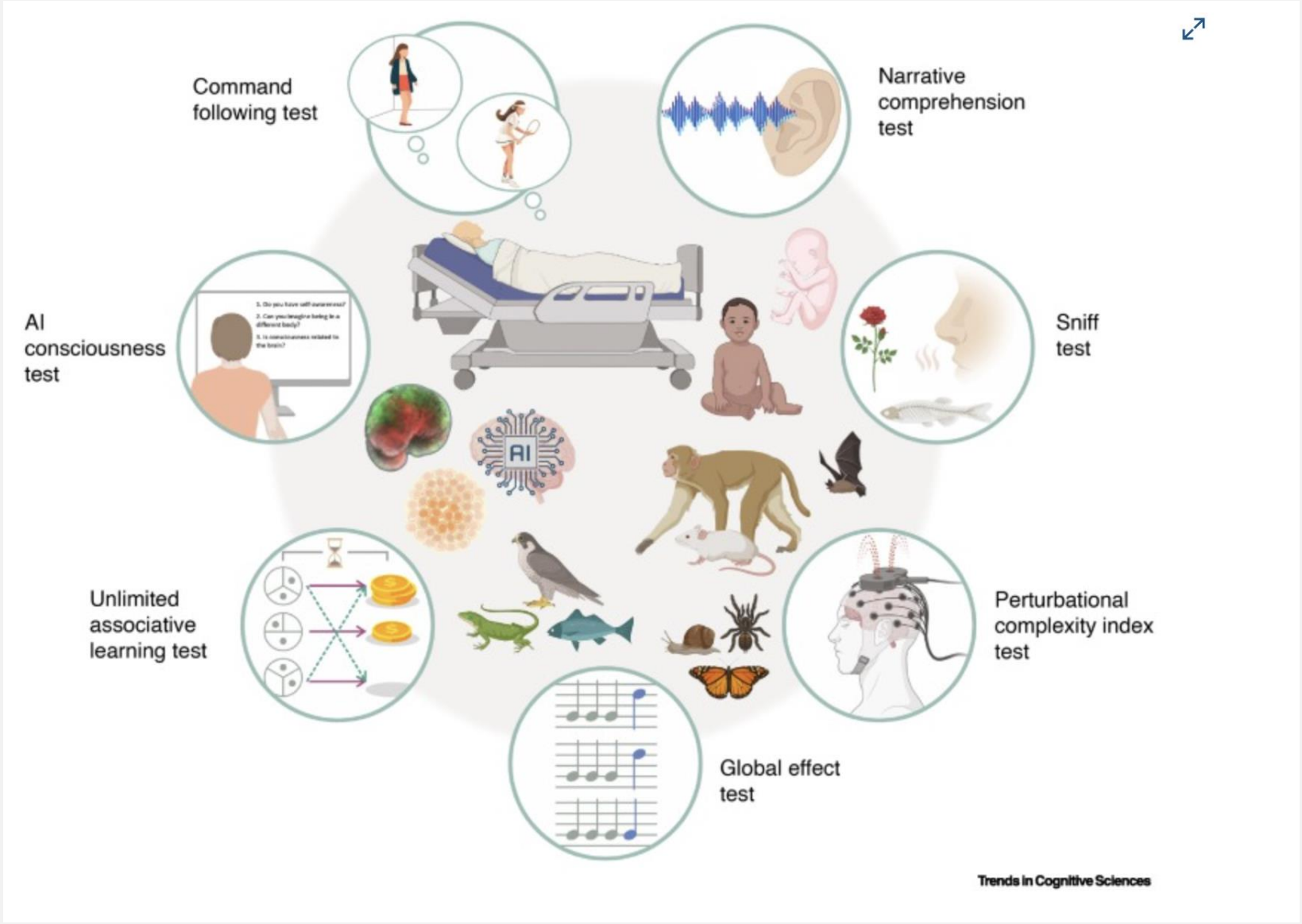
THEORY NEUTRAL

- *Problèmes :*

- *Ces approches ne semblent pas être neutres par rapport à la théorie, elles entrent en conflit avec elle, en laissant entendre que les détails architecturaux sur lesquels les différentes théories se concentrent n'ont pas d'importance - ou qu'ils sont incomplets (cf. Shevlin : la tâche de dire quelles capacités cognitives sont pertinentes n'est-elle pas une tâche théorique ?)*

THEORY NEUTRAL

- *Bayne et al:*


Command following test
Narrative comprehension test
AI consciousness test
Sniff test
Unlimited associative learning test
Perturbational complexity index test
Global effect test
AI
Trends in Cognitive Sciences

		Command following	Narrative	Sniffing	PCI	Global effect	ACT	UAL
ALTERED STATES	Sedation	+	+	+	+	+	?	+
	Epileptic seizure	+	+	+	+	+	?	+
	Sleep/Dreaming	+ ?	+ ?	+ ?	+	+ ?	?	+ ?
UNCLEAR CAPACITY FOR CONSCIOUSNESS	Disorders of consciousness	+	+	+	+	+	?	+
	Babies	−	?	+	+	+	−	+
	Fetuses	−	−	?	+	+	−	?
NON-HUMAN ANIMALS	Non-human mammals	−	−	+	+ ?	+	−	+
	Non-mammal vertebrates	−	−	−	−	+	−	+
	Invertebrates	−	−	−	−	?	−	+
ARTIFICIAL SYSTEMS	Neural organoids	−	−	−	?	+ ?	−	+ ?
	xenobots	−	−	−	?	+ ?	−	+ ?
	AI	+ ?	+ ?	−	?	+ ?	+	+ ?

Trends in Cognitive Sciences

Figure 2 The scope of tests for consciousness (C-tests).

THEORY HEAVY

THEORY HEAVY

- Deux versions :
- 1) S'entendre sur la théorie finale, l'appliquer.
- 2) Choisis une distribution de probabilités, applique-la.

THEORY HEAVY

- Le problème avec le premier : plus facile à dire qu'à faire !
- Problème avec le second : les résultats ne seront pas concluants (cf. le rapport Butlin).

THEORY HEAVY

- Par ailleurs, bien que le rapport Butlin décrive notre approche de cette manière, elle peut également être interprétée comme une théorie légère.

THEORY HEAVY

- Problème majeur : le problème de l'exigence *(demandingness)* / de la spécificité

- Compte tenu de notre théorie préférée pour le cas humain, dans quelle mesure l'appliquons-nous « strictement » à d'autres cas ?

THEORY HEAVY

- La plupart des théories de la conscience ont été conçues en tenant compte du cas humain (et peuvent ne viser qu'à distinguer les états conscients des états inconscients, en supposant que l'entité en question est capable de conscience) :
- Elles identifient une structure importante dans le contexte de la biologie humaine et de l'architecture neuronale ....
- Mais si l'on soustrait cette toile de fond, l'élément clé est minimal et peut être instancié dans des systèmes très petits (trop petits).

THEORY HEAVY

- Critère éxigeante:

- Trop éxigeante, compte ténu nos raisons pour accepter ces theories en premier lieu (on presuppose que les sujets humains sont conscients…)

THEORY HEAVY

- Critère permissive / libérale:

THEORY HEAVY

- **Théorie de l'espace de travail global**:

- *Un nœud centralisé qui transmet à une majorité d'autres nœuds.*

THEORY HEAVY

- **Théorie de la pensée d'ordre supérieur :**

- *Un nœud qui suit/réfléchit le contenu d'un autre nœud*

THEORY HEAVY

- **Théorie de la mémoire à court terme fragile**:

- *Contenu représenté sur plusieurs couches d'un réseau récurrent.*

THEORY HEAVY

- **Théorie de la réafférence / du tronc cérébral**:

- *Un nœud qui détermine si les changements apportés à un autre nœud sont endogènes ou exogènes.*

THEORY HEAVY

- **Critére indéterminée (incrementalism, rejectionism)**

THEORY HEAVY

- dire que dans les cas difficiles, il n'y a pas de vérité ni faussété (c’est **vague**).
- Cf, combien de cheveux est le seuil pour être chauve ?
- combien de dollars est le seuil pour être riche?
- combien de grains de sable est le seuil pour qu'il s'agisse d'un tas?

THEORY HEAVY

- Mais peut-on vraiment donner un sens à cette idée ? Et que signifie-t-elle dans la pratique ? Crée-t-elle plus de problèmes qu'elle n'en résout ?

THEORY LIGHT

THEORY LIGHT

- Birch suggère que, pour le cas de la conscience animale en tout cas, nous pouvons identifier des marqueurs de conscience (des caractéristiques partagées par tous les systèmes dont nous convenons qu'ils sont conscients), puis nous pouvons extrapoler que d'autres systèmes que nous découvrons avec ces marqueurs sont également conscients.

THEORY LIGHT

- Birch propose des capacités spécifiques telles que le **conditionnement par traces** (*trace conditioning*), une forme d'apprentissage par conditionnement où il y a un délai entre le stimulus et la récompense.

THEORY LIGHT

- Aussi: **l'apprentissage par inversion** (*reversal learning)* et **l'apprentissage multisensorial** (*multisensory learning)*

THEORY LIGHT

- Problème : cela peut fonctionner pour les créatures avec lesquelles nous partageons une lignée évolutive, mais c'est moins évident lorsqu'il s'agit de systèmes artificiels (pour lesquels, par exemple, le conditionnement des traces est trivial).

THEORY LIGHT

- Dex problèmes ici:
- 1) Les indicateurs en question sont tout simplement trop faibles, et donc inadaptés au cas de l'IA
- (*mais pourquoi la faiblesse est-elle une forme d'inadéquation ?*).

THEORY LIGHT

- **Conditionnement des traces**:
- **Seuil**: Le conditionnement par traces nécessite la capacité de garder un stimulus en mémoire au fil du temps, ce qui est lié à des structures cérébrales plus avancées telles que l'hippocampe. Il a été observé chez les **mammifères**, les **oiseaux** et certains **insectes**.
- **Présente chez**: De nombreux mammifères (rats, chiens, primates, etc.), oiseaux (pigeons, etc.) et certains insectes (abeilles, etc.) peuvent effectuer un conditionnement par traces. La fonction hippocampique est essentielle chez ces espèces.
- **Absent chez**: Les animaux dont le système nerveux est plus simple, comme la **plupart des invertébrés** (vers, mollusques, etc.), n'ont généralement pas la complexité neurologique nécessaire pour effectuer un conditionnement de traces, car leur système de mémoire est moins développé.

THEORY LIGHT

- **Analogue de l'IA**: Dans l'apprentissage par renforcement (RL), en particulier dans les algorithmes qui utilisent des réseaux neuronaux récurrents (RNN) ou des transformateurs avec mémoire, les systèmes peuvent apprendre à associer des stimuli à des récompenses différées. Par exemple, les méthodes d'apprentissage **par** renforcement dotées de représentations d'état suffisantes peuvent apprendre à gérer des tâches où la récompense est différée, de la même manière que les animaux apprennent à associer un stimulus à un résultat après un certain laps de temps.
- **Limites**: Bien que le RL puisse limiter la tâche de conditionnement de traces, le conditionnement de traces biologique implique des mécanismes de mémoire spécifiques, tels que les fonctions hippocampiques, alors que les systèmes d'IA gèrent cette tâche par le biais d'unités de mémoire abstraites (RNN, couches d'attention) qui ne reproduisent pas la nature nuancée, adaptative et continue de la mémoire animale.
-

THEORY LIGHT

- **<u>Apprentissage par inversion</u>**:
- **Seuil**: L'apprentissage par inversion dépend de la flexibilité cognitive, en particulier de la capacité à inhiber les associations précédemment apprises et à en former de nouvelles. Il apparaît chez les espèces dont le cortex préfrontal ou les structures analogues sont plus développés.
- **Présent chez**: La plupart des **mammifères** (primates, rongeurs, etc.), des **oiseaux** (corbeaux, pigeons, etc.) et même certains **reptiles** et **poissons** (poissons rouges, etc.) peuvent pratiquer l'apprentissage inversé.
- **Absent chez**: Les espèces dont le système nerveux est moins complexe, comme de nombreux **invertébrés** et **amphibiens**. Ces animaux éprouvent souvent des difficultés à s'adapter lorsqu'une association apprise est inversée, ce qui témoigne d'une flexibilité cognitive moindre.

THEORY LIGHT

- **Apprentissage par inversion dans l'IA**:
- **Analogue de l'IA**: Les **algorithmes d'apprentissage par renforcement**, en particulier ceux qui utilisent les méthodes d'apprentissage Q ou de gradient de politique, peuvent mettre en œuvre l'**apprentissage par inversion** en ajustant leurs politiques lorsque la fonction de récompense change. Essentiellement, lorsqu'un comportement appris n'est plus gratifiant, le système met à jour sa politique pour rechercher de nouveaux comportements qui donnent lieu à des récompenses positives.
- **Limites**: Si les systèmes RL peuvent en principe gérer l'apprentissage par inversion, leur flexibilité et leur rapidité d'adaptation aux nouvelles structures de récompense dépendent de facteurs tels que les taux d'apprentissage, les stratégies d'exploration et l'architecture du modèle. Les organismes biologiques ont tendance à faire preuve d'une capacité d'adaptation plus robuste et plus efficace en raison de leurs systèmes de prise de décision plus sophistiqués, en particulier dans des environnements complexes et non stationnaires.

THEORY LIGHT

- **Apprentissage multisensoriel**:
- **Seuil**: La capacité d'intégrer des informations à travers les modalités sensorielles nécessite un cerveau plus avancé avec des régions dédiées au traitement et à l'intégration des différentes entrées sensorielles.
- **Présentes chez**: L'apprentissage multisensoriel est observé chez de nombreux **vertébrés** (mammifères, oiseaux, reptiles) et certains **insectes** (par exemple, les abeilles) qui naviguent dans des environnements complexes. **Les primates** et les **cétacés** présentent une intégration multisensorielle particulièrement avancée.
- **Absent chez**: De nombreux **invertébrés** et animaux plus simples (par exemple, les **vers plats**, les **éponges**) n'ont pas cette capacité, car leur traitement sensoriel est plus isolé ou manque de coordination centrale pour l'intégration des différents sens.

THEORY LIGHT

- **Apprentissage multisensoriel dans l'IA**:
- **AI Analogue**: Les **systèmes multimodaux**, tels que les transformateurs texte-image (par exemple, CLIP, DALL-E), intègrent déjà plusieurs modalités sensorielles (par exemple, texte et image) pour générer des prédictions, des classifications ou des résultats. Ces systèmes utilisent un espace d'intégration partagé pour combiner des informations provenant de différentes modalités, de la même manière que les animaux intègrent les données provenant de différents sens.
- **Limites**: Les systèmes d'IA excellent souvent dans l'intégration multisensorielle dans des contextes spécifiques et prédéfinis (par exemple, la conversion de texte en image, le sous-titrage d'images). Cependant, contrairement aux systèmes biologiques, ils s'appuient généralement sur de grandes quantités de données d'entraînement et sur des tâches prédéfinies, et ne disposent pas de la capacité d'adaptation et de généralisation dynamique et en temps réel dont font preuve les animaux lorsqu'ils traitent des stimuli multisensoriels dans des environnements nouveaux.

THEORY LIGHT

- Deux problèmes ici:
- 2) Il peut être facile de « jouer » les marqueurs que nous identifions. (« *the gaming problem* »)
- Shevlin, Dung et Birch-Andrews examinent le problème des jeux, la loi de Goodhart, etc. ..... Il est possible de créer un système qui passe le test (systèmes ad hoc de Dung).

THEORY LIGHT

- le principe « même comportement - même cause » peut avoir un sens pour une lignée évolutive commune, mais pas tellement pour des systèmes dont nous savons qu'ils sont totalement différents.
- -Exemple : la parole...
- -Pour nous, le système est construit sur l'incarnation, la perception sensorielle multimodale et énactive, la mémoire épisodique, etc.
- -Pour le GPT4, il se contente de mémoriser des modèles à partir des mots de l'internet...

THEORY LIGHT

- Dung suggère:
- Nous avons juste besoin d'une norme pour les cas où un modèle/système est conçu de manière ad hoc.

THEORY LIGHT

- Mais la conscience d'un système ne devrait-elle pas être intrinsèque, et ne pas dépendre de la raison pour laquelle il a été créé ? (Un indice que quelque chose ne va pas dans la solution de Dung)

THEORY LIGHT

- Syndrome ou maladie : il s'agit également d'une question plus générale. Notre objectif est-il d'identifier d'autres indicateurs (une description plus complète du syndrome) ou considérons-nous la conscience comme une « maladie » (cf. l'hépatite ... ou l'hépatite C) ?

THEORY LIGHT

- Que recherchons-nous ici ? Quels sont les types de nature (*natural kinds)* qui peuvent être en jeu ?
- Existe-t-il des types naturels ***de calcul*** ?

THEORY LIGHT

- Platon (Phèdre 265e) :
- Nos meilleures théories « *découperont la nature au niveau des articulations* »
- --- Depuis lors, les philosophes débattent de ce que cela signifie