COMMENT SAVOIR SI LES ROBOTS SONT CONSCIENTS ?

La question de trouver une méthodologie appropriée pour la découverte

Jonathan Simon

PHI-6385

Séance 6

APERCU

APERCU

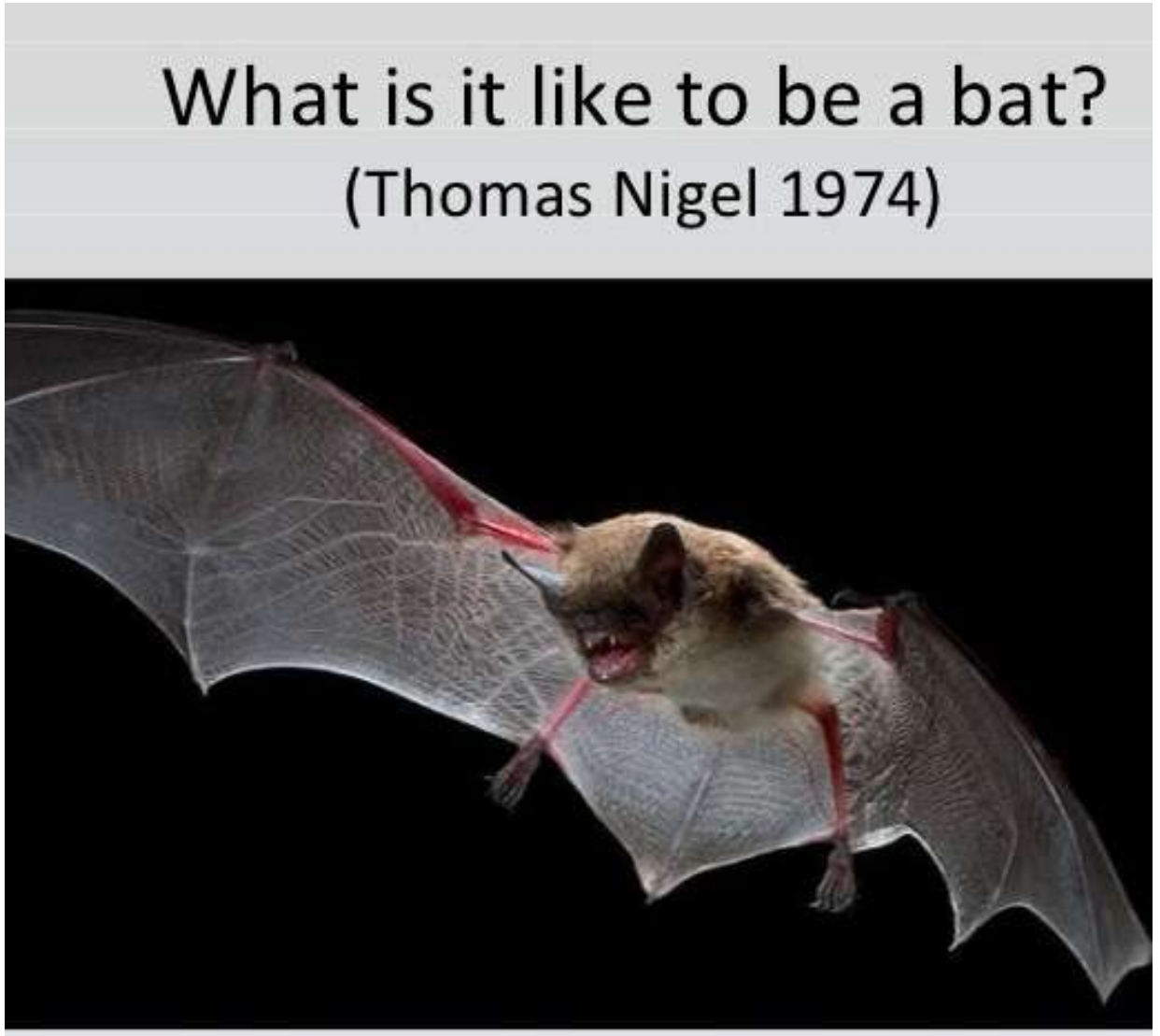
- 1) La question de la méthodologie
- 2) Méthodologie neutre sur le plan théorique «Theory-neutral methodology» (Schneider, Shevlin, Tye)
- 3) Méthodologie legère sur le plan théorique «Theory-light methodology»
- (Birch, Andrews)
- 4) Méthodologie lourde sur le plan théorique «Theory-heavy methodology» (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 legère sur le plan théorique «Theory-light methodology»
- 3) Méthodologie lourde sur le plan théorique «Theory-heavy methodology»

METHODOLOGIE

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

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 legère sur le plan théorique «Theory-light methodology»

Appel à des principes théoriques largement acceptés, par exemple le principe selon lequel l'histoire évolutive commune des mécanismes signifie généralement une similarité de fonction.

METHODOLOGIE

- 3) 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.

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)...*

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

- *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 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, une forme d'apprentissage par conditionnement où il y a un délai entre le stimulus et la récompense.

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

- Il peut être facile de "jouer" les marqueurs que nous identifions.

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).