Conscience et Intelligence artificielle(s)

Vues par Jacques Pitrat

Par l'intermédiaire
de Gérard Sabah

Essai de définition(s) (Robert)

- **Usage courant**
  - *être éveillé (se rendre compte)*
  - *connaissance immédiate spontanée*
- **Conscience psychologique**
  - *connaissance de ses propres connaissances*
  - *perception de ses opérations mentales (introspection)*
- **Conscience morale**
  - *faculté de juger ses propres actes*
- **(+ conscience professionnelle, liberté de conscience…)**

Limites de l'intelligence humaine

- **Lenteur des neurones (et nombre limité)**
- **Impossibilité de modifier la structure du cerveau**
- **Mémoire de travail très limitée**
- **Capacité d'expliquer ce qu'on fait (réflexivité) mais incapacité de méta-expliquer (pourquoi on a ces intuitions)**
- **Inaccessibilité de l'inconscient**

Pourquoi s'intéresser à la conscience ?

- **Edelman :** ***Les fonctionnalités nécessaires à une véritable intelligence sont celles qui, fondées sur l'inconscient, permettent l'émergence de la conscience chez l'homme***

- ***Pourquoi pas chez les machines ?***

Convergences

- **Résolution de problèmes**
  - → ***conscience réflexive + conscience morale***
- **Traitement automatique des langues**
  - → ***inconscient + conscience réflexive***

Quelques idées de base (J. P.)

- **Les chercheurs en IA ont deux défauts : *trop intelligents et pas assez paresseux***
- **L'IA est le problème le plus difficile auquel l'homme s'est attaqué**
- **L'homme n'est peut-être pas assez intelligent pour le résoudre**
- **Il faut s'aider des systèmes d'IA eux-mêmes**
- **➔ amorçage**

Idées de base…

- **META : Après résolution pb, généraliser et apprendre**
- **On ne copie pas l'intelligence humaine mais il est souvent bon de s'en inspirer**
- **Cognition artificielle ≠ cognition humaine**
  - ***Certaines capacités cognitives artificielles sont inaccessibles aux humains (et vice versa)***
- **N'oublions pas l'IA forte !** ***(AFIA 100^{e})***

Conscience artificielle

- **Conscience réflexive**
  - ***Apprendre (s'observer)***
  - ***S'adapter (se modifier)***
- **Conscience morale**
  - ***Autonomie***
  - ***Choix***
- **Certains aspects réalisés dans CAIA**
  - ***Système général de résolution de problèmes donnés sous forme de contraintes***
- **Décrits dans *Artificial Beings : The Conscience of a Conscious Machine (ISTE & Wiley, 2009)***

CAIA (Chercheur Artificiel en Intelligence Artificielle)

- **Déclarativité et réification**
  - ***Modification simple par changement de la valeur d'une variable***
- **Pile des appels de fonctions**
  - ***Arrêts sans problèmes***
  - ***Détection d'anomalies***
- **Modification dynamique**
  - ***Création et compilation de nouveaux programmes***

Avantages des systèmes artificiels « conscients »

- **Analyser ce qu'on sait**
- **S'observer en train de fonctionner**
- **Méta-combinatoire**
- **Immortalité**

+ *éventuellement* meilleure compréhension de la conscience humaine

Analyser ce qu'on sait

- **Accès à toutes les connaissances (≠*humains*)**
- **Compréhension**
- **Forme déclarative**
  - ***Analyse, création, modification plus faciles***
- **Transformées sous forme procédurale**
  - ***Utilisation de méta-connaissances***
  - ***450 000 lignes de C (10 000 règles conditionnelles)***
  - ***Efficacité***

S'observer en train de fonctionner

- **Différence fondamentale entre humains et machines : l'inconscient et la mémoire**
  - ***On connaît certaines étapes de nos raisonnements, mais on ne sait pas <u>pourquoi</u> on y a pensé***
  - ***CAIA peut rendre tout « conscient »***
- **Interruptions et reprises aisées**
- **Explication, méta-explication**

Méta-combinatoire

- **Chaque méthode de CAIA est associée à**
  - ***des déclencheurs potentiels***
  - ***des conditions qui peuvent l'interdire***
  - ***des priorités qui déterminent quand l'utiliser***
- **➔ Combinatoire sur les méthodes elles-mêmes (+ explication)**
- **Exemple :**
  - *Trouver tous les nombres m et n positifs et inférieurs à 10^{18} tels que:*
    $4*m + 3*n^2 = 817\,401\,078\,957\,542\,034$

Immortalité

- **Facilité de reproduire un système**
- **Copies identiques ou légèrement différentes**
  - ***Prise de risques***
  - ***Test de divers variantes***
  - ***Adaptation au problème***

Conscience morale

- **Autonomie**
  - ***Capacité de s'enrichir de sa propre expérience***
  - ***Modèle réflexif nécessaire***
- **Respect des valeurs et des principes**
  - ***Évolutifs chez l'homme (3 niveaux)***
  - ***Fixes (et ≠) chez les machines***
- **Impossibilité de prévoir tous les comportements d'un programme d'IA**
- **Nécessité d'une surveillance robuste**
  - ***Validation et contrôle ?***

Conclusion

- **Énorme potentiel de l'IA**
  - ***Implémentation et expérimentation nécessaires***
- **Deux problèmes pour l'IA forte**
  - ***Intelligence humaine***
  - ***Structure de la recherche***
    - Temps consacré
    - Long terme
- **Espoirs : sciences cognitives et amorçage mais, *une perspective à 100 ans…***

MERCI !