3e colloque des doctorants et jeunes chercheurs
*Philosophie, Langages & Cognition*
6 juin 2011

# Des collaborations possibles entre Intelligence Artificielle et philosophie de l'esprit

**Robin Lamarche-Perrin**

**Laboratoire d'Informatique de Grenoble**
**Université Joseph Fourier**
**Université Pierre-Mendès-France**

# Introduction

- IA et *philosophie de l'IA*

  → Une science et une « méta-science »

- IA et *philosophie de l'esprit*

  → Deux disciplines au même niveau

- « L'IA comme *science philosophique*. »
  [Andler, 1984], préface trad. de [Dreyfus, 1979]

# Deux hypothèses [Searle, 1999]

## IA FAIBLE

- Problème pratique
  - *Les machines peuvent **simuler** des comportements intelligents*

- **Spécialistes de l'IA**
  - Conférence de Dartmouth

« Every feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. » [McCarthy *et al*., 1955]

  - Simulation cognitive [Minsky]

## IA FORTE

- Problème philosophique
  - *Les machines sont **réellement** intelligentes (*e.g.*, états mentaux, intentionnalité, conscience).*

- **Philosophie de l'esprit**
  - *Brain replacement* [Glymour]
  - Chambre chinoise [Searle, 1980]

# Trois approches

**IA faible → IA forte**

- Le Behaviourisme logique [Turing, 1950]

**IA faible ↔ IA forte**

- Le computationnalisme [Newell & Simon, 1976]

**IA faible ← IA forte**

- La critique de Dreyfus [Dreyfus, 1979]

# Le Behaviourisme et le computationnalisme

AI faible → AI forte  &  AI faible ↔ AI forte

# Le Test de Turing [Turing, 1950]

- IA faible (simuler l'intelligence)
  - Une définition comportementale de l'intelligence
  - Behaviourisme méthodologique
- IA forte (engendrer la conscience)
  - « Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks. » [Turing, 1950]
  - Behaviourisme logique
- IA faible → IA forte
  - Convention : le comportement intelligent induit nécessairement la conscience
  - **Élimination de la philosophie de l'esprit !**

# Le Computationnalisme

## [Newell & Simon, 1976]

- *The Physical Symbol System Hypothesis*
  - « A physical symbol system has the necessary and sufficient means for general intelligent action. » [Newell & Simon, 1976]
  - « *sufficient* »
    → Les ordinateurs peuvent agir intelligemment. (IA faible)
  - « *necessary* »
    → Les cerveaux sont des systèmes symboliques.
    → Les cerveaux et les ordinateurs sont similaires.
    → Les ordinateurs peuvent avoir des états mentaux. (IA forte)
- IA faible ↔ IA forte
  - IA et philosophie travaillent sur une même catégorie d'objets

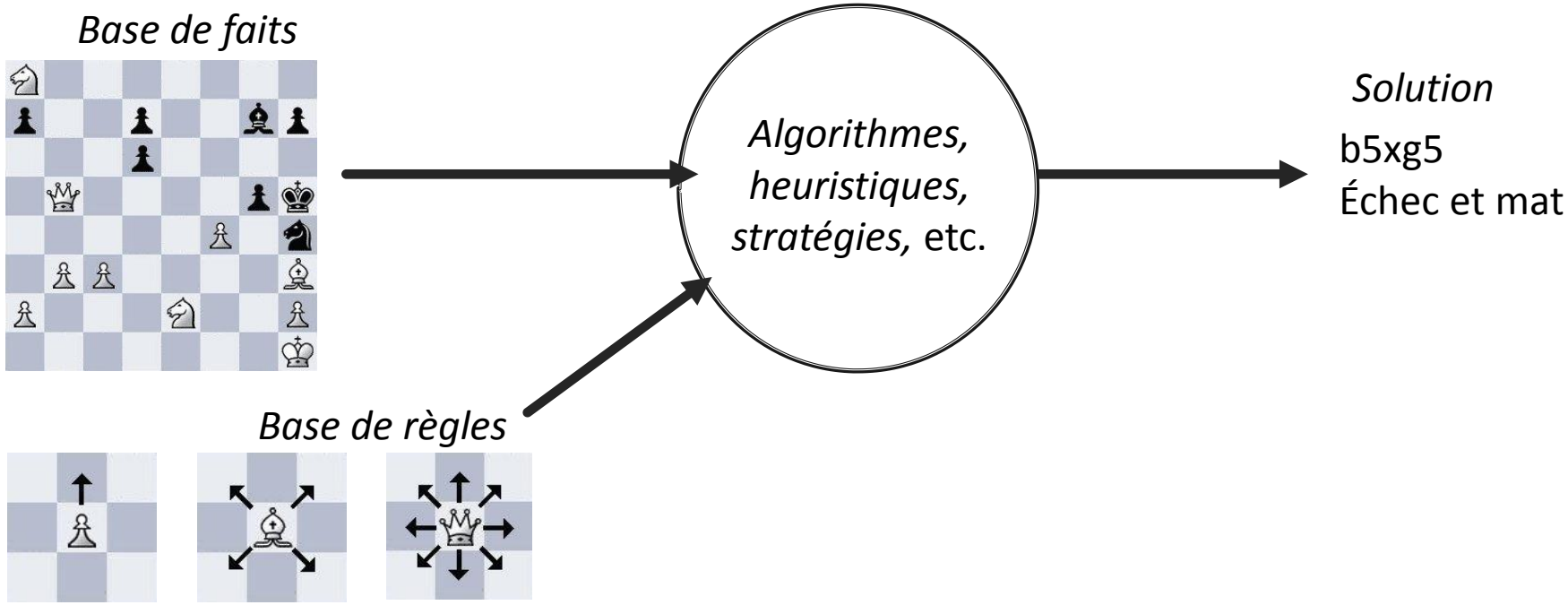
# Les Critiques de l'Intelligence Artificielle

# La Critique de Searle [Searle, 1980]

- Critique de l'IA
  - La chambre chinoise [Searle, 1980]
  - La syntaxe, la sémantique et l'intentionnalité
- La critique de Searle
  - Est mal renseignée sur les recherches en IA
  - Est une objection de principe à l'IA
  - Atteint uniquement le behaviourisme et le « computationnalisme linéaire »
  - **S'intéresse à l'IA forte seulement !**

# La Critique de Dreyfus [Dreyfus, 1979]

- Partie 1 : bilan critique
  - Échec du *General Problem Solver* [Newell & Simon, 1963]
  - Échec du paradigme computationnaliste



# La Critique de Dreyfus [Dreyfus, 1979]

- Partie 2 : trois erreurs philosophiques
  - Présupposé psychologique (computationnalisme)
  - Présupposé épistémologique
  - Présupposé ontologique

- Le poids de la tradition
  - Hobbes, Descartes, Leibniz, Kant, le 1er Wittgenstein
  - L'étouffement du connexionnisme [Dreyfus & Dreyfus, 1988]

# La Critique de Dreyfus [Dreyfus, 1979]

- Partie 3 : dépasser la tradition
  - La phénoménologie, le 2$^{d}$ Wittgenstein, Heidegger
  - *Knowing-that* et *knowing-how*
    [Dreyfus & Dreyfus, 1986]

- La critique de Dreyfus
  - Est informée sur les recherches de l'IA
  - N'est pas une opposition de principe à l'IA
  - Cible le paradigme dominant
  - **S'intéresse à l'IA faible !**

# Une véritable collaboration

AI faible ← AI forte

# De la critique à la pratique

- IA forte → IA faible
  - Une théorie de l'esprit adéquate

  → de bons résultats en pratique

- L'héritage de Dreyfus et de la phénoménologie
  - Connexionnisme, IA distribuée
  - Robotique évolutionniste, vie artificielle
  - Anti-représentationnalisme, systèmes dynamiques
  - Tournant pragmatique, robotique incarnée, énaction

# Un exemple de collaboration

[Lamarche-Perrin, 2011]

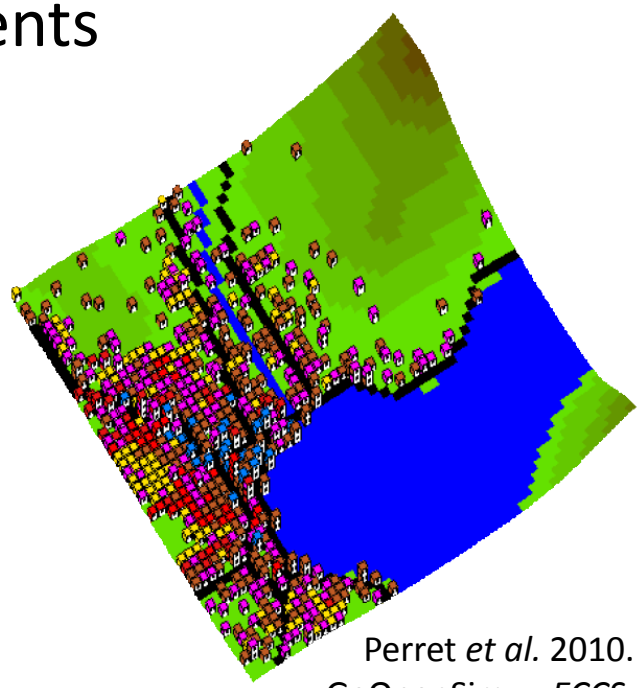
- Simulation des phénomènes émergents
  - Exemple : les dynamiques urbaines

**Microscopiques**

*Dynamique des individus*

**Macroscopiques**

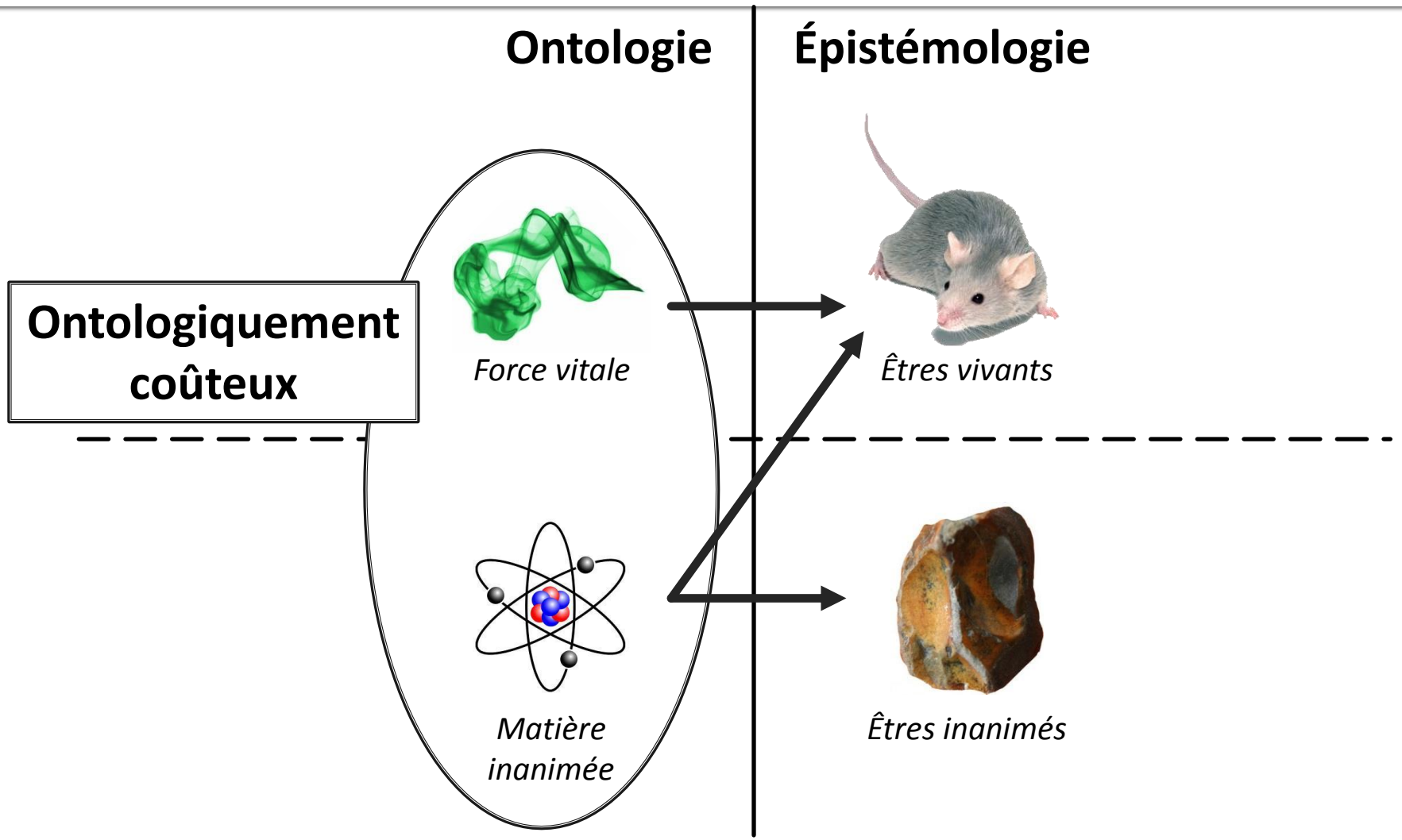
*Marché immobilier*



Perret *et al.* 2010. « GeOpenSim ». *ECCS*.

- Des concepts philosophiques…
  - Philosophie britannique [Mill, Broad]
  - Dualisme, monisme, éliminativisme, émergence épistémique
- …aux simulations informatiques

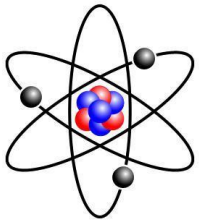
# Le Dualisme non-réductionniste



# Le Monisme éliminatif

**Ontologie** | **Épistémologie**

**Épistémologiquement faible**



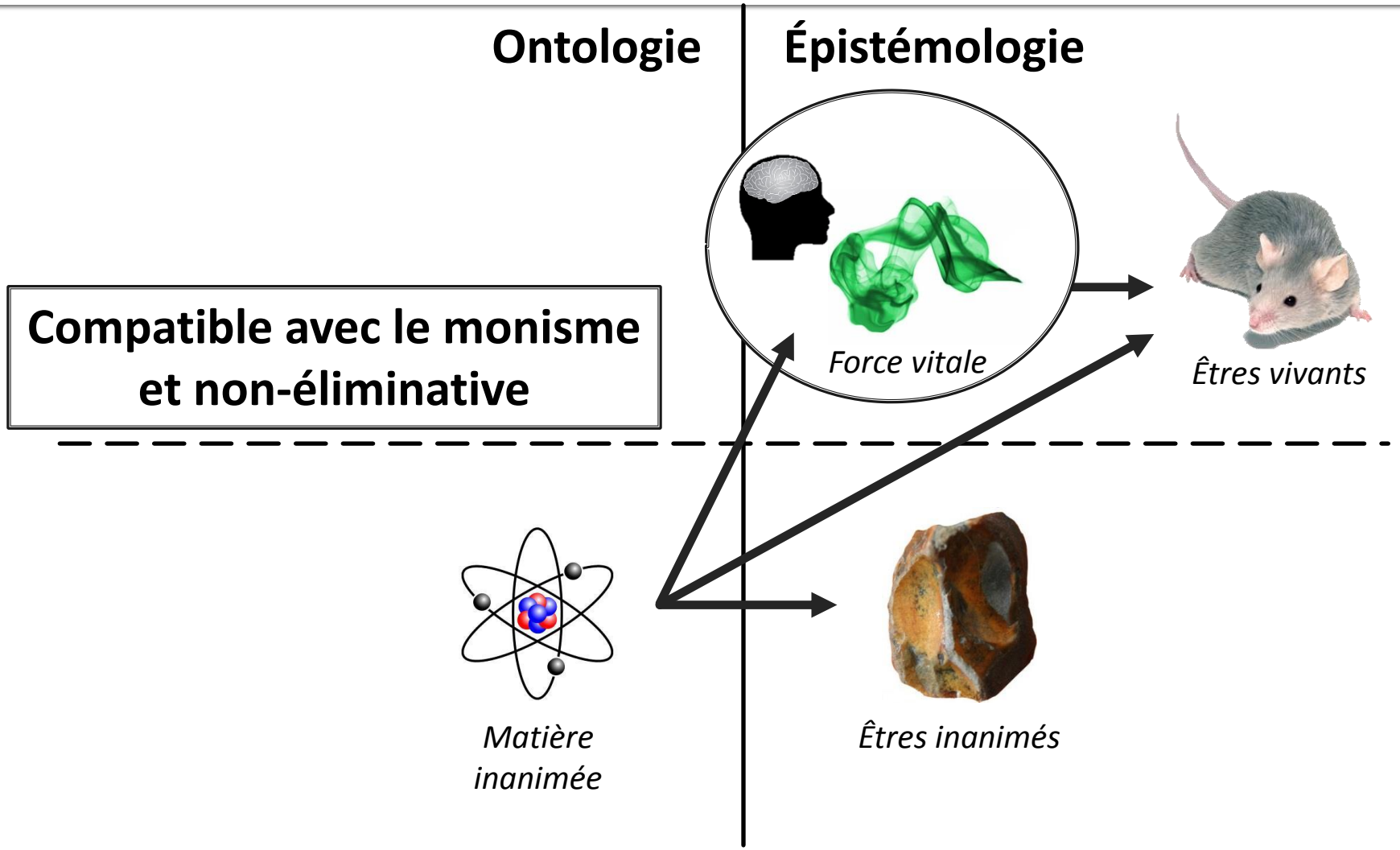
*Matière inanimée*

&

*Êtres inanimés*

*Êtres vivants*

# L'Émergence épistémique



# Appliquer l'émergence épistémique

- Analogie
  - Ontologie → Conception des systèmes
  - Épistémologie → Analyse des systèmes
- Deux contraintes méthodologiques
  - Monisme microscopique
  - Non-éliminativisme

# Le Dualisme (analogie)

**Conception** | **Analyse**

Niveau macroscopique


*Marché immobilier*

Niveau microscopique

*Dynamique des individus*

# Le Monisme microscopique

**Conception** | **Analyse**

Niveau macroscopique


*Marché immobilier*

Niveau microscopique

*Dynamique des individus*


*Marché immobilier*

- Véritable approche *bottom-up*
- Conception *épiphénoméniste*

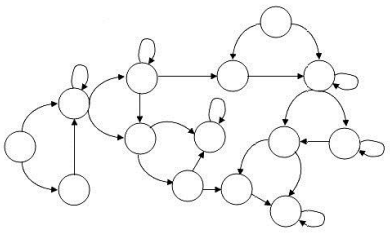
# Le Monisme éliminatif (analogie)

**Conception** | **Analyse**

Niveau macroscopique

Niveau microscopique

*Dynamique des individus*



*Description microscopique*

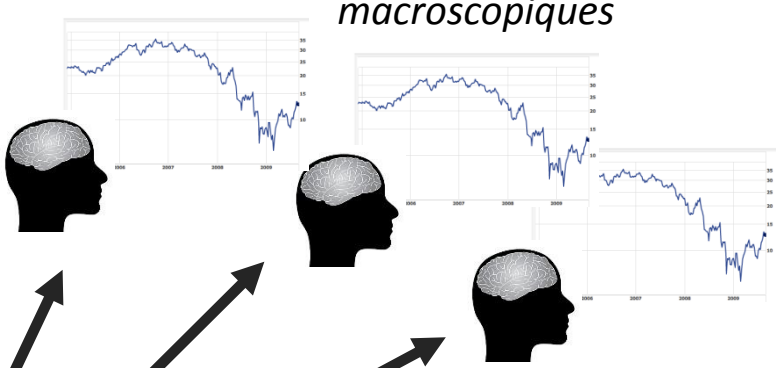
# Le Non-éliminativisme

**Conception**

- Points de vue multiples
- Conception *pragmatique*

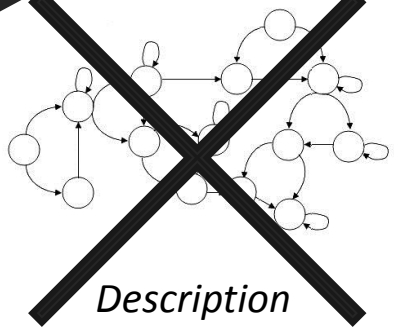
**Analyse**

*Descriptions macroscopiques*



Niveau microscopique

*Dynamique des individus*



*Description microscopique*

# Critique des conceptualisations en IA

- Approches dualistes
  - Systèmes à « tableaux noirs »[Sawyer, 2001]
  - Systèmes multi-modèles [Gil-Quijano *et al*., 2010]
- Approches monistes
  - Auto-organisation [Picard, 2004]
  - **Émergence par détection [Bonabeau & Dessalles, 1997]**
- Approche éliminatives
  - Compréhension et simulation [Darley, 1994]
  - Coûts computationnels [Bedau, 1997]
- Approche non-éliminatives
  - **Émergence par détection [Bonabeau & Dessalles, 1997]**

# De la pratique à la critique

- Non IA faible → Non IA forte
  (échec de l'IA → échec de la philosophie)
  - Le computationnalisme a échoué

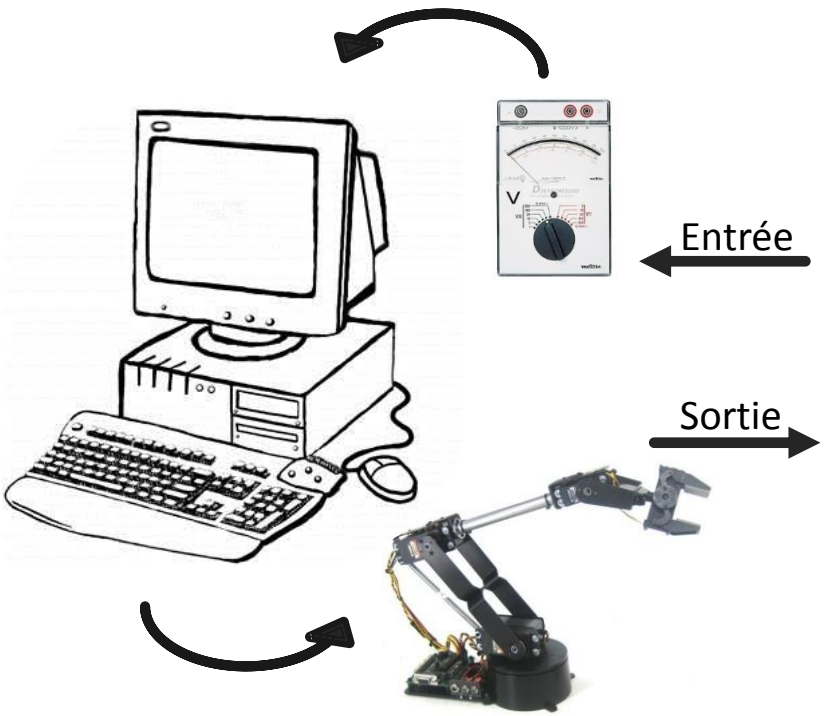
  → son cadre philosophique est remis en cause

- L'IA comme banc d'essais de la philosophie
  - Falsifier et évaluer les théories de l'esprit
  - Une *philosophie expérimentale* [Harvey, 2000]
  - Une *science philosophique* [Andler, 1984]

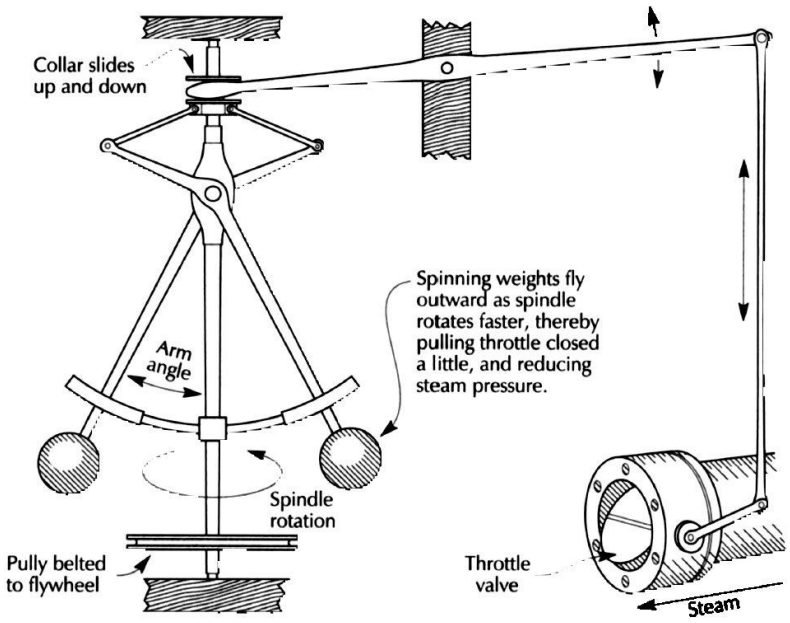
# Deux types de régulateurs

## [Van Gelder, 1996]

### RÉGULATEUR SYMBOLIQUE




### RÉGULATEUR DYNAMIQUE




**The Watt centrifugal governor**

Adapted by [Van Gelder, 1996]
from [Farey, 1827]

# L'esprit est quel type de régulateur ?

## RÉGULATEUR SYMBOLIQUE

- Computationnalisme
  - Utilise des représentations
  - Indépendant de l'environnement
  - Temps discret
- Opérations logiques

```
1  public class TestPetitLien extends TestCase {
2      public void testPetitLien() throws IOException {
3          LittleLinkRequest request =
4              new LittleLinkRequest("MonUrlTresLong", "Alias");
5          LittleLinkRequest request2 =
6              new LittleLinkRequest("MonUrlTresLong", "4");
7          try {
8              String petitLien = request.getLittleLink();
9              String petitLien2 = request2.getLittleLink();
10         } catch (LittleLinkException e) {
11             e.printStackTrace();
12 }   }   }
```

## RÉGULATEUR DYNAMIQUE

- Systèmes dynamiques
  - N'utilise pas de représentations
  - Couplé avec l'environnement
  - Temps continu
- Équations différentielles

$$\frac{d^2\theta}{dt^2} = (n\omega)^2 \cos\theta \sin\theta - \frac{g}{l}\sin\theta - r\frac{d\theta}{dt}$$

# Conclusion

# Synthèse

- La philosophie au service de l'IA

  → De « bonnes » théories de l'esprit font de « bonnes » applications.

- L'IA au service de la philosophie

  → Un banc d'essai pour évaluer les théories de l'esprit.

- L'IA, la philosophie de l'esprit et l'éthique

  → Quels critères pour la conscience ?

# Un problème éthique

- Et pour l'IA forte ?
  - Critère comportemental [Turing]
  - Critère fonctionnel [Fodor]
  - Critère biologique [Searle]
  - Critère corporel [Dreyfus]
  - Critère quantique [Penrose]
  - *Etc.*

# Merci pour votre attention

# Bibliographie


- Dreyfus, L.H. 1979. *What Computers Can't Do*. New York: MIT Press. (Préface de Andler, D.)
- Dreyfus, L.H. et Dreyfus, S.E. 1986. *Mind Over Machine : The Power of Human Intuition and Expertise in the Era of the Computer*. New York: The Free Press.
- Dreyfus, L.H. et Dreyfus, S.E. 1988. « Making a Mind versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint ». *Artificial Intelligence*, vol. 117, n°1, p. 15–43.
- Harvey, I. 2000. « Robotics: Philosophy of Mind Using a Screwdriver ». *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol. III, p. 207–230.
- Lamarche-Perrin, R. 2011. « Conceptualisation de l'émergence : dynamiques microscopiques et analyse macroscopique des SMA » *Plate-forme AFIA 2001 : FUTURAMA*.
- McCarthy, J., Minsky, M.L., Rochester, N. et Shannon, C.E. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*.
- Newell, A. et Simon, H. 1963. « GPS, A Program that Simulates Human Thought. » Dans Feigenbaum, E.A. et Feldman, J. (eds.) *Computers and Thought*, New York: McGraw-Hill.
- Newell, A. et Simon, H.A. 1976. « Computer Science as Empirical Inquiry: Symbols and Search ». *Communications of the ACM*, vol. 19, n°3, p. 113–126.
- Searle, J. 1980. « Minds, Brains and Programs ». *Behavioral and Brain Sciences*, vol. 3, n°3, p. 417–457.
- Searle, J. 1999. *Mind, Language and Society*. New York: Basic Books.
- Turing, A. 1950. « Computing Machinery and Intelligence ». *Mind*, vol. LIX, n°236, p. 433–460.
- Van Gelder, T. 1996. « Dynamics and cognition ». Dans Haugeland, J. (éd.) *Mind Design II*, Bradford/MITP.