VALEUR NUMÉRIQUE

PHI 6385

Séance 13

PROGRAMME

- 1) Les fondements du statut moral (Sentience vs Agence)
- 2) Les énigmes concernant le statut moral de l'IA
- 3) Implications pratiques du statut moral des IA

STATUT MORAL

STATUT MORAL

- Patience morale : être une chose telle que ce qui t'arrive est (intrinsequement) moralement important.
- Agence morale : être une chose qui peut agir, et donc porter la responsabilité de ces actions.

FONDEMENTS DE STATUT MORAL

- Deux grandes conceptions des fondements d'être un patient moral:
- Sentientisme et Agentivité

FONDEMENTS DE STATUT MORAL

- Sentientisme:
- Être un patient, c'est être conscient ou subir des états conscients affectifs (comme le plaisir, la douleur, l'émotion, le désir).

FONDEMENTS DE STATUT MORAL

- Sentientisme:

- (compare : la conscience est ce qui fait de toi un sujet moral vs... la conscience te donne des intérêts spécifiques, per exemple d'éviter de ressentir de la douleur, qui tendent à avoir plus d'importance que ceux des agents non conscients).

FONDEMENTS DE STATUT MORAL

- Agentivisme :
- être un patient moral, c'est être un agent (moral).

- Kagan

FONDEMENTS DE STATUT MORAL

- Les implications pour l'idée que les IA auraient / n'auraient pas de statut moral, contribuent-elles elles-mêmes à façonner nos intuitions sur ce qu'il faut pour être conscient / un agent....?

FONDEMENTS DE STATUT MORAL

- Si une théorie de la conscience semble trop exclusive ou trop inclusive sur le plan moral, cela peut-il justifier de la réévaluer ?

ENIGMES

ENIGMES

- 1) Le problème des implémentations minimales :
- Dans la plupart des théories de la conscience ou de l'agentivité, il existe des seuils subtils, et on peut trouver des choses qui semblent similaires (qui semblent avoir des intérêts similaires) mais qui ne satisfont pas la théorie. Surtout avec l'IA où l'on peut bricoler les modèles....

ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :

- Pour un agent de type RL - qu'est-ce qu'il trouve gratifiant ?

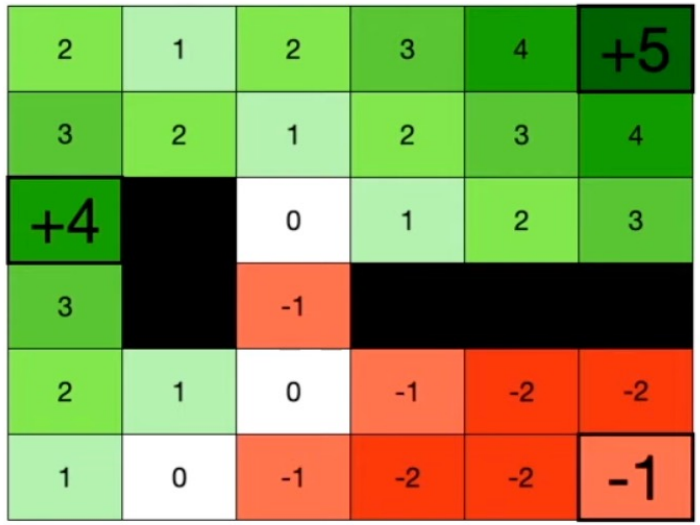
ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :

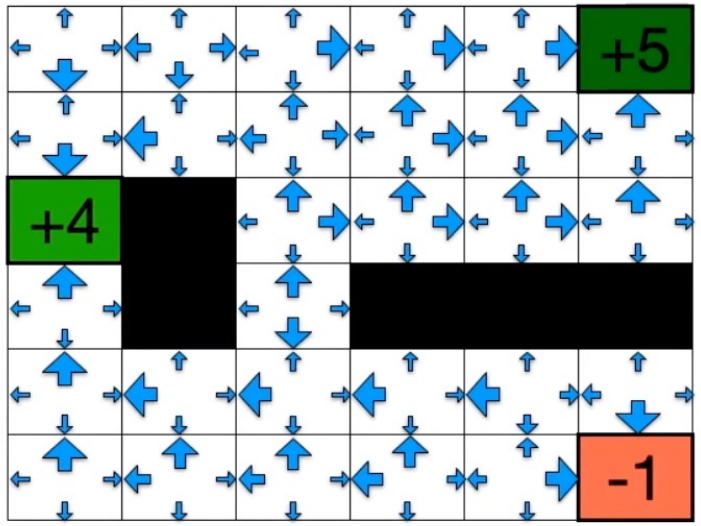
- Y a-t-il une différence entre une récompense et l'absence de punition ?

Two neural networks

Value neural network



Policy neural network



Luis Serrano

ENIGMES

- 2) Le problème de la détermination de ce que seraient réellement les intérêts des IA. :

- Pour un "agent" LLM, veut-il ce qu'il "prétend" vouloir ?


Joining for coffee at a cafe
KM:
AC:
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?
Finishing a morning routine
JM:

IMPLICATIONS

IMPLICATIONS

- 1) Personnalité juridique
- 2) Couts bizarres
- 3) Risques bizarres
- 4) Évite de créer des IA qui ont un statut moral (Metzinger), ou évite simplement celles pour lesquelles il n'est pas clair qu'elles aient un statut moral (Schwitzgebel et Garza)?

PERSONNALITÉ JURIDIQUE

- Peut-on leur donner une personnalité juridique ? Est-ce une bonne idée ? Leurs créateurs pourraient-ils s'en servir pour se décharger de leur responsabilité (par exemple, c'est la voiture qui l'a fait, pas moi !)?

COUTS BIZARRE

- 1) bénéficiaires ordinaires
- 2) bénéficiaires «super»

COUTS BIZARRE

- 1) bénéficiaires ordinaires

- Ont-ils des droits contre l'esclavage ? Des droits au confort ? La poursuite du bonheur ? En quoi cela dépend-il de ce qu'ils désirent ? S'ils ont des droits à la survie, cela signifie-t-il que nous devons sauver leurs poids modèles, même si cela devient très coûteux ?

COUTS BIZARRE

- 2) bénéficiaires super: Shulman et Bostrom

BOSTROM ET SHULMAN

- La possibilité des **super-bénéficiaires**

BOSTROM ET SHULMAN

- La possibilité des **super-bénéficiaires**

- Une entité qui tirerait beaucoup plus d'utilité d'un bien, ou beaucoup plus de désutilité d'un mal, que les humains.

BOSTROM ET SHULMAN

- Les super-bénéficiaires sont-ils possibles ?

- Shulman et Bostrom: passent en revue un certain nombre de raisons - différentes façons dont les super-bénéficiaires pourraient se produire

BOSTROM ET SHULMAN

- Possibilité 1) ce sont peut-être des êtres comme nous, mais ils sont plus nombreux que nous.
- Possibilité 2) ils sont comme nous à chaque moment subjectif, mais ils ont plus de moments subjectifs
- Possibilité 3) nous les programmons pour qu'ils soient plus faciles à satisfaire

BOSTROM ET SHULMAN

- Ce n'est pas seulement une version de la conclusion répugnante de Parfit : là, le souci est qu'un nombre bien plus grand de vies médiocres semblerait (pour l'utilitariste) être meilleur qu'un nombre plus petit de bonnes vies.
- En revanche, ici, un scénario à envisager est que ce sont nos vies qui semblent médiocres par comparaison

BOSTROM ET SHULMAN

- De plus, bien qu'il y ait de nombreux points communs avec les questions d'éthique animale, ici, il peut être plus difficile pour les déontologues d'affirmer que les humains ont un statut plus élevé que l'autre type d'esprit en question : on peut dire que les animaux sont des agents inférieurs, il est plus difficile de voir comment cela pourrait être le cas pour les IA...

RISQUES BIZARRES

- . .. Note le conflit potentiel avec la sécurité des IA : si elles ont le droit de ne pas être fermées, mais que nous prévoyons qu'elles deviennent dangereuses....

EVITER DE LES CRÉER

- Metzinger vs Schwitzgebel et Garza vs e/acc

- Est-ce possible?
- Est-ce trop tard?