Pensée artificielle ?

Séance 12, Philosophie de l'esprit H2020

Jonathan Simon

Plan

- Intro : Qu'est-ce que l'IA ?
- Turing: Oui, Pourquoi pas? (le jeu de l'imitation)
- Searle: Non -- une syntaxe sans sémantique ne suffit pas (la salle chinois)
- Dreyfus: Non -- le savoir-faire quotidien n'est pas basé sur la représentation

Intro : Qu'est-ce que l'IA ?

La question

- Un système d'IA peut-il être intelligent (on laisse la conscience pour plus tard)

- 1) Qu'est-ce que l'intelligence?

- 2) Comportement / performance de l'intelligence
- -- La majeure partie de la littérature porte sur ce point : que faudrait-il pour qu'un système d'IA se comporte de manière intelligente (passer le test de Turing, faire preuve de bon sens, tirer la bonne morale d'une histoire, s'adapter aux circonstances pour atteindre ses objectifs.... assez ouvert)

L'IA classique

- Systèmes de traitement de symboles (GOFAI)
- Considérez les états d'une machine de Turing comme spécifiant / représentant directement les pensées, les plans, les instructions au niveau humain
- L'IA précise beaucoup de faits et de règles formelles : en effet, le genre de théorie de tout dans la réalité quotidienne que Descartes voulait déduire des premiers principes
- Inspiré par Turing, il est devenu un projet de recherche sérieux au milieu des années 50. S'est épanouie dans les années 60 et 70.

Connectionnisme

- L'approche rivale, née dans les années 60 et 70.
- Réseaux de neurones : structure informatique modelée sur les réseaux de neurones biologiques
- L'unité de calcul de base est modelée sur la mise à feu d'un neurone (beaucoup d'opérations très simples et subsymboliques - pas de représentation explicite des faits ou des règles au niveau humain, au lieu de cela, ceux-ci sont codés implicitement dans des poids neuronaux)

Connectionnisme

- Notez que les réseaux neuronaux peuvent être mis en œuvre sur des machines de type Turing (et même avec des processeurs GPU / TPU plus efficaces, pas vraiment un processeur séparé pour chaque neurone).

- Les neurones eux-mêmes sont représentés mathématiquement - tout n'est qu'un tas d'algèbre linéaire (matrices, vecteurs, tenseurs)

Connectionnisme

- J'insiste sur ce point parce que beaucoup de gens parlent de la différence entre l'IA symbolique (classique) et l'IA subsymbolique (connexionniste) comme s'il s'agissait d'une différence entre un ordinateur qui procède comme le décrit Turing, et un ordinateur fait d'un tas de petites puces neuronales qui ont une fonctionnalité très limitée. Ce n'est pas le cas :par exemple, votre ordinateur portable en implémente un si vous utilisez la fonction d'écran de fond de Zoom.

Connectionnisme

- Le point important est que les états de machines internes pertinents pour l'IA classique sont des états qui codent explicitement des faits et des principes sur le monde, écrits sous forme symbolique, et les opérations de calcul formelles sont censées refléter le raisonnement (ainsi, le raisonnement est une question de manipulation purement syntaxique des symboles)

Connectionnisme

- En revanche, pour les réseaux de neurones, les états internes sont des grappes de matrices (tenseurs) qui représentent des groupes de neurones (une matrice représente une couche, une ligne de la matrice représente un neurone), et les opérations (généralement la multiplication des matrices) représentent des événements de tir neuronal : des événements où chaque neurone transmet un signal unique aux neurones auxquels il est directement connecté dans la couche suivante

Comparer les deux:

- Trier les chats des chiens :
- Supposons que nous ayons des images en pixels que nous voulons alimenter à un système : soit une image de chat, soit une image de chien
- GOFAI:Il faut d'abord des règles pour convertir les motifs de pixels en bords, puis les motifs de bords en formes, puis les motifs de formes en, par exemple, oreille de chien, nez de chat, etc. A la dernière étape, vous avez terminé. Mais essayez de penser à une règle que vous pourriez écrire explicitement, pour déduire quand une forme est en forme d'oreille de chien plutôt que de chat...

Comparer les deux:

- Trier les chats des chiens :
- Supposons que nous ayons des images en pixels que nous voulons alimenter à un système : soit une image de chat, soit une image de chien
- Réseau neuronal : vous l'entraînez sur beaucoup, beaucoup d'images (pour lesquelles vous lui dites si c'est un chien ou un chat). Ensuite, il ajuste ses poids (les valeurs exactes dans les matrices), en ajustant une frontière très complexe (côté chien, côté chat) dans un espace de représentation de valeurs de pixels. Ensuite, lorsque vous lui montrez une nouvelle image, il repère simplement de quel côté de la limite il se trouve.

Comparer les deux:

- (Ne vous inquiétez pas si cela n'a pas vraiment de sens. Le fait est qu'il n'y a pas de règles explicitement formulées, juste une correspondance abstraite de motifs)

Comparer les deux:

- Ainsi, pour contextualiser, Turing envisage ce que nous appellerions aujourd'hui une IA classique

- Les objections de Searle et Dreyfus ne sont pas fondées sur un quelconque aperçu des technicités informatiques, mais les deux peuvent être considérées comme complémentaires aux critiques connectionnistes. (Bien que Dreyfus anticipe également certains défis auxquels le connexionnisme doit encore faire face)

Turing: Machines Informatique et Intelligence

Alan Turing

Turing

- Turing est surtout connu pour ses travaux sur la théorie de la calculabilité : il a identifié le problème de l'arrêt (die Entscheidungsproblem), a fourni des définitions concrètes de nombreux termes centraux de la théorie du calcul (comme celui d'un ordinateur universel, alias une machine de Turing). Il a également travaillé à la décodage de la machine Enigma (le système de cryptographie nazi) pendant la Deuxième Guerre mondiale.
- En parallèle, il a rédigé ce document, fixant l'agenda de la recherche sur l'IA pour les 50 prochaines années

Machines Informatique et Intelligence

- 1) Les machines peuvent-elles penser ?
- propose Turing : il est trop difficile de répondre à cette question. Remplacez-la par :
- 2) Une machine peut-elle être performante (tromper quelqu'un 30 % du temps) au jeu d'imitation ?

Machines Informatique et Intelligence

- Le jeu de l'imitation :

- Un examinateur, deux joueurs. Un joueur est humain, l'autre est une machine. Les joueurs sont dans des salles séparées, seuls des messages écrits sont échangés. L'examinateur peut demander n'importe quoi à l'un ou l'autre des participants, le but étant de deviner qui est l'humain. L'objet du jeu est de tromper l'examinateur en lui faisant croire que l'autre joueur est la machine.

Machines Informatique et Intelligence

- Pourquoi propose-t-il ce remplacement ?
- 1) Il évite un débat philosophique insolubles, le remplace par une question qui reste assez substantielle et difficile à répondre
- 2) Comportementalisme / Positivisme ?
- 3) Théorie cartésienne de la primauté de la preuve de la capacité linguistique (rappel de Descartes vs. Romanes et Huxley)

Machines Informatique et Intelligence

- Le reste du document :
- 1) la formulation de ce que doit être un ordinateur numérique (une machine de Turing) et de ce que doit être l'exécution d'un programme informatique (une machine à états discrets).

- 2) Ensuite, les réponses aux objections prévues pour affirmer qu'un ordinateur numérique sera capable de passer le test 50 ans plus tard (en 2001)

- 3) Enfin une esquisse positive d'une voie à suivre : ici, quelques réflexions classiques sur l'IA. Notez cependant qu'il ne les utilise pas pour répondre à des objections...

Les machines de Turing

- Nous en avons déjà parlé, dans la classe sur le fonctionnalisme.

- L'idée principale : Une machine à états discrets est spécifiée par une table de machine, une table de commandes comme : si vous êtes à l'état S1, et que vous avez l'entrée I1, allez à l'état S3 et sortez O2...
- Une machine de Turing est une machine qui peut mettre en œuvre n'importe quel programme de ce type, étant donné que ses entrées, ses sorties et ses états sont formellement codables (par exemple dans les opérations arithmétiques)

Les machines de Turing

- Le vrai génie de tout cela est qu'il est possible de démontrer que pratiquement toute opération formelle (qui intuitivement est quelque chose qui peut être fait sans nécessiter de perspicacité ou d'imagination, quelque chose que "n'importe quel ordinateur pourrait faire") peut être encodée arithmétiquement de la manière appropriée

Réussir le test

- Quel est exactement son argument selon lequel une telle machine pourrait passer le test ? Il n'a pas vraiment d'argument décisif : il examine plutôt les objections et y répond.

- À noter ici : en particulier, il ne fonde pas explicitement tout sur l'idée que toute pensée humaine est une manipulation formelle de symboles : son objectif principal est de faire valoir qu'un ordinateur numérique pourrait réussir le test, et non qu'une "IA classique" pourrait. Les hypothèses qui conduisent à l'IA classique n'arrivent qu'à la fin

Objections

- Objection théologique
- La tête dans le sable
- Objection mathématique
- Argument de la conscience
- Divers handicaps
- L'objection de Lady Lovelace
- La continuité du système nerveux
- L'informalité du comportement
- Perception extra-sensorielle

Objections

- Objection théologique: Nous ne pouvons pas créer des âmes ! Turing : en effet. Mais ici, la question est seulement de savoir si nous pouvons créer des manoirs pour eux.
- La tête dans le sable: C'est effrayant ! Turing : pas d'objection
- Objection mathématique: Limitations formelles sur ce que les ordinateurs peuvent penser / prouver ! Turing : pourquoi de telles limitations ne peuvent-elles pas exister aussi pour les connaissances humaines ?

Objections

- Argument de la conscience: Une machine ne peut jamais vraiment être consciente ? Turing : vous pourriez changer d'avis si vous en voyiez une passer le test. Nous pouvons également aborder la question de savoir si la machine peut passer le test sans adresser celui-ci
- Divers handicaps: Mais une machine ne sera jamais capable de... Turing : soit non pertinent (manger de la glace), soit suppliant (réussir ce test), soit faux (faire des erreurs)
- L'objection de Lady Lovelace: Les machines ne peuvent pas vraiment créer / engendrer quoi que ce soit. Turing : dans un sens pertinent, il n'est pas évident que nous non plus.

Objections

- La continuité du système nerveux: Le système nerveux est continu et non discret. Turing : l'examinateur en jeu ne pourrait pas en tirer profit.

- (q : oui, mais cela pourrait-il avoir une importance pour que la machine soit suffisamment performante ?)

Objections

- L'informalité du comportement: Le comportement humain ne peut pas être décrit par un système de règles !

- Turing : Peut-être qu'elles peuvent être son juste difficile. Notez combien il serait difficile de déduire les règles qui régissent un ordinateur réel simplement en observant son comportement...
- Perception extra-sensorielle: ?

Apprentissage machine

- Turing conclut son article par quelques réflexions sur la façon d'amener les machines à apprendre à bien jouer (plutôt que de les programmer explicitement).

- Ses idées ici anticipent certaines idées dans l'apprentissage du renforcement, mais vont aussi clairement dans le sens de l'IA classique. Mais il ne s'appuie pas sur ces idées dans la première partie de l'article pour défendre ses hypothèses clés...

Searle: la salle chinoise

La salle chinois

- Pourquoi faut-il toujours parler de la Chine (la nation chinoise, la chambre chinoise) ? Je ne sais pas.

- L'argument de Searle : même si un ordinateur numérique passe le test de Turing (et est capable de converser couramment), il ne comprend toujours pas ce qu'il dit (que cela suffise ou non pour l'"intelligence" est une autre question, mais il y a vraisemblablement un lien)

La salle chinois

- Voici comment vous pourriez faire partie d'une machine de Turing qui parle chinois (même si, supposons, vous ne le faites pas. Remplacez le chinois par toute langue que vous ne parlez pas du tout):
- Version simplifiée (courriel). Supposons que je parle chinois. Si A vous envoie un message en chinois, vous me le transmettez. Je rédige une réponse, vous la transmettez à A. Nous faisons cela toute la journée. Vous parliez chinois ?

La salle chinois

- Il est clair que non !
- Maintenant, la version de Searle. Au lieu d'un courriel, vous êtes dans une boîte, et quelqu'un à l'extérieur de la boîte vous remet un message en chinois, avec un petit numéro dans le coin. Vous avez un grand livre qui vous montre, avec le numéro dans le coin, ce qu'il faut dessiner sur une page, puis vous le distribuez à l'extérieur de la boîte.
- Il s'avère que le livre d'instructions que vous avez en main correspond exactement à ce que je vous aurais dit de dire. Mais en ce qui vous concerne, vous ne faites que chercher des chiffres, puis vous écrivez des gribouillis.

La salle chinois

- Maintenant, vous faites vraiment partie d'un ordinateur numérique qui met en œuvre un programme de conversation chinoise. Mais il est clair que vous ne parlez pas chinois. Donc aucun autre ordinateur ne met en œuvre ce programme non plus.

La salle chinois

- Une réponse courante : d'accord, peut-être que vous ne parlez pas chinois, mais vous n'êtes qu'une partie de l'ordinateur numérique. Peut-être que le système dans son ensemble parle chinois ?

- La réponse de Searle, en effet : mais au-delà de vous, ce n'est qu'une boîte, un livre et quelques cartes et marqueurs. Si vous ne parlez pas chinois, votre système et celui des autres ne parlent pas chinois non plus.

La salle chinois

- Autre réponse populaire : le système doit être relié au monde par une relation de cause à effet : entrées perceptuelles et sorties motrices

- Searle : peut-être, mais ce n'est pas suffisant : on peut facilement ajouter cela. Disons qu'en plus du gribouillage - les sorties, il y a d'autres marques que vous pouvez produire qui feront que certains appendices se déplaceront à l'extérieur de votre boîte. Et dites qu'en plus des entrées de gribouillis, vous obtenez des photos sous forme d'images, de ce qui est à l'extérieur de la boîte. Vous ne parlez toujours pas chinois !

La salle chinois

- Quelle est la morale de l'histoire pour Searle ?
- Pas claire !
- La syntaxe ne suffit pas pour la sémantique.
- La biologie est nécessaire à la conscience.
- Un programme pourrait-il suffire, mais pas celui-ci ? Ou a-t-il montré qu'aucun programme informatique ne pouvait suffire ? Si oui, comment le simple fait de la biologie peut-il aider ?

Dreyfus: Ce que les ordinateurs ne peuvent (encore) pas faire

Hubert Dreyfus

- dans la tradition de Heidegger et Merleau-Ponty
- A commencé à critiquer l'IA classique bien avant qu'elle ne soit populaire (les années 60 et 70)
- Cette deuxième édition (avec sa nouvelle introduction) a été écrite en 1992, à la fin de la première vague d'IA classique, début de la première montée du connectisme.
- Beaucoup de ses critiques préfigurent (ou même motivent) le connectionnisme, mais il critique aussi le connectionnisme (et ses critiques des deux écoles restent valables aujourd'hui !)

Hubert Dreyfus

- Critiques primaires :

- l'intelligence humaine est une question de savoir-faire pratique profond, et ce savoir-faire pratique n'est pas équivalent à une quelconque quantité de connaissances-que

Hubert Dreyfus

- Si un expert doit expliquer au système ce qui est pertinent / approprié quand, le système ne sera jamais pleinement compétent / adaptable

- Lorsque nous apprenons quelque chose de nouveau, comme les échecs, nous appliquons des règles qu'un expert nous a expliquées. Mais à mesure que nous acquérons de l'expérience, nous apprenons à improviser, à adapter et à contourner les règles. Pour les vrais experts, la bonne chose à faire est simplement évidente, ce n'est pas une question de raisonnement...