UNIVERSITÉ LAVAL

L’IA est-elle dangereuse ? La réponse est oui à moins que …

Brahim Chaib-draa

12 menaces qui pourraient détruire le monde d'ici 100 ans

- Les menaces qui pèsent sur l'humanité ne manquent pas à en croire un groupe de chercheurs de la fondation des défis globaux et de l'université d'Oxford.

les 12 menaces

Cette présentation vise à …

- IA est-elle dangereuse pour l'humain ?
  - Dans combien de temps ?
  - Faut-il agir ?
  - Si oui, quoi faire ?
  - Comment le faire ?
- Cette présentation tente de répondre à ces questions.
- Stola et Yampolskly (2015) [1].

UNIVERSITÉ LAVAL

Plan de la présentation

- IA faible
  - « Définition »
  - Ses avantages
  - Ses Dangers
- IA forte : AGI
  - Qu'est ce que l'IA forte ?
  - Est-elle possible ?
  - Ses risques pour l'humanité ?
- Comment contrer les risques de AGI
  - Propositions pour actions venant de la société
  - Propositions pour des contraintes externes sur AGI
  - Recommandations pour une conception d'AGI safe

IA faible

- Dans l'IA faible, l'ordinateur digital est juste vu comme un outil pour étudier l'intelligence et développer une technologie utile.
- Un programme IA est tout au plus une simulation d'un processus cognitif, mais lui-même n'est pas un processus cognitif.
  - Ex: une simulation de voiture ne constitue pas une voiture.

L'IA faible : avantages

- Avantages via des innovations technologiques
  - Traitement de la parole, Vision, Perception
  - Applis pour la mobilité
  - Jeux, interfaces, et autres
  - Sécurité (ift, routière, etc.)
  - Robotique
  - Biologie
  - Santé
  - Etc.

L’IA faible : avantages

- Avantages pour des tâches
  - Répétitives;
  - Non-sécuritaires ou dangereuses
  - Pénibles
  - Trop complexes pour des humains
  - Etc.

L'IA faible : dangers

- Possibles dangers d'un IA autonome
  - Vers la guerre avec 0 pertes
    - Drones
    - Autres systèmes d'armements autonomes
- Possibles dangers technologiques
  - IA + Big Data + Google's Quantum Computer
  - IA + Cloud + Google's Quantum Computer
  - IA + Nano Technologie
- Dangers d'une automatisation accrue :
  - Perte de jobs

UNIVERSITÉ LAVAL

L'IA forte ou AGI

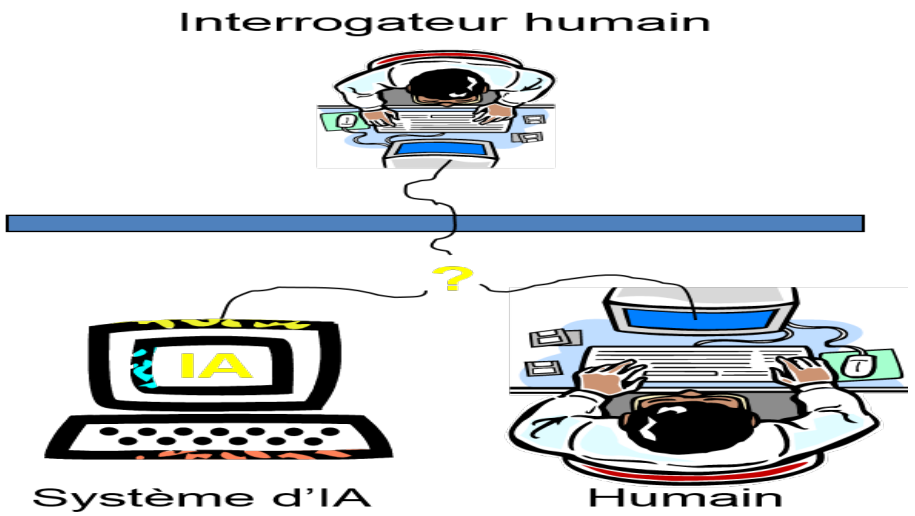
- L'IA forte, aussi appelée AGI pour Artificial General Intelligence, suppose qu'un ordinateur peut être programmé pour être lui-même un processus conscient (a mind)
- Mind (Wikip.) /ˈmaɪnd/ is the set of cognitive faculties that enables consciousness, perception, thinking, judgement, and memory—a characteristic of humans, but which also may apply to other life forms.[3][4]

AGI comme Mind

- AGI pourrait être
  - Intelligente
  - Comprendre
  - Percevoir
  - Avoir des croyances/connaissances
  - Exhiber d'autres états cognitifs généralement attribués au « humains ».

Tests pour confirmer que AGI est opérationnelle

- Test de Turing



- The Coffee Test (Goertzel) : une machine est sensée effectuée la tâche qui consiste à aller dans une maison d'Américain moyen et trouver comment faire le café.

UNIVERSITÉ LAVAL

Tests pour AGI est opérationnelle (2)

- The Robot College Student Test (Goertzel) : une machine est sensée se faire recruter dans une université, prendre des cours et passer des examens (comme le ferait un humain ou mieux ?) et obtenir un diplôme.
- The Employement Test (Nilsson) : une machine est sensée travailler en faisant un job (selon les lois du marché) et performer aussi bien, sinon mieux qu'un humain faisant le même travail.

AGI est-elle possible ?

- 4 niveaux de Penrose [2].
  - A. Toute pensée se réduit à un calcul; en particulier, le sentiment de connaissance immédiate consciente naît simplement de l'exécution de calculs appropriés (AGI ou IA forte).
  - B. La connaissance immédiate est un produit de l'activité physique du cerveau; mais bien que toute action physique puisse être simulée par un calcul, une telle simulation ne peut par elle-même susciter la connaissance immédiate (IA faible).

Les 4 niveaux de Penrose

- C. La connaissance immédiate est suscitée par une action physique du cerveau, mais aucun calcul ne peut simuler cette action physique (Penrose).
- D. On ne peut expliquer la connaissance immédiate à l'aide du langage de la physique, de l'informatique, ni de quelque autre discipline scientifique que ce soit.
  - Ce niveau nie résolument la possibilité de progresser dans la compréhension de la conscience.

Les thèses de Penrose (suite)

- On ne peut ni démontrée ni réfutée, aucune des thèses A, B, C.
- Le débat n'est pas tranché, toutefois bien des chercheurs croient à un possible avènement de AGI dans les [20 à 100] prochaines années.
  - Signature d'une pétition au dernier AAAI;
  - E. Musk; S. Hawking, B. Gates et bien d'autres ont sonné l'alarme
  - E. Musk a mis sur la table 10 millions pour une IA sécure.
  - OpenAI (10 billions $)

UNIVERSITÉ LAVAL

OpenAI

- OpenAI is a non-profit artificial intelligence research company. Its goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return.
- Since our research is free from financial obligations, we can better focus on a positive human impact. We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as possible.

Les risques de l'AGI

- L'AGI permettra d'automatiser la plupart des tâches;
- L'AGI risque de porter préjudice à l'humain;
- L'AGI peut devenir puissante rapidement (alors qu'on n'est pas encore préparé à ça).

UNIVERSITÉ LAVAL

AGI automatisera les tâches

- AGI remplacera les travailleurs pour des raisons de coût, d'efficacité et de qualité.
- Pour certaines tâches, AGI fera mieux en peu de temps et avec peu d'erreurs
- AGI pourrait apprendre à faire différents types de tâches, ou même n'importe quel type de tâche, sans nécessiter un effort de re-engineering
- Comme de plus en plus de tâches seront automatisées, l'IA faible pourrait s'avérer insuffisante et on aurait besoin de plus en plus d'AGI.

AGI pourrait porter préjudice à l'humain

- Des outils mathématiques très sophistiqués ont contribué à la crise financière de 2010.
- Des systèmes automatiques de défense ont parfois engendré des catastrophes touchant leur propre utilisateurs.
- Comme les machines vont devenir de plus en plus autonomes et de plus en plus intelligentes; les humains auront de moins en moins d'opportunités d'intervenir, ceci mènera à de plus en plus de machines.

AGI risque de porter préjudice à l'humain (2)

- ….de plus en plus de machines; alors risques si
  - AGI n'est pas robuste;
  - AGI a des objectifs non alignés avec ceux des humains
  - AGI n'a pas d'étique. Etc.
- Si AGI devraient être de plus en plus puissante et indifférente aux humains : les conséquence peuvent être désastreuses.

AGI est son préjudice à l'humain (3)

- Yudkowsky a dit : 'The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else'.
- Les AGI doivent-ils suivre un comportement que dicte l'économie et la théorie des jeux
  - Maximiser son utilité selon le principe de la rationalité
  - Sinon vous êtes vulnérables à d'autres agents qui pourraient vous exploiter.

AGI est son préjudice à l'humain (4)

- Les agents qui suivent ce principe pourraient être tenté de se reproduire
- La multiplication des agents pourraient les amener à vouloir acquérir des ressources sans tenir compte des autres
- Ces mêmes agents pourraient être amenés à s'auto-améliorer et à persister à réaliser leurs buts.
- Concevoir des AGI qui pourraient respecter les valeurs humaines ? Pas évident du tout.

Les risques de l'AGI

- L'AGI peut devenir puissante très vite; Ceci peut être causé par :
  - Un dépassement du hardware
    - GPU—DNN (Chip)---hardware de plus en sophistiqué
  - Une explosion de la vitesse
    - Google's machine is making the leap from 512 qubits to more than a 1000 qubits.
  - Une explosion de l'intelligence.
    - AGI + **humain**; AGI + **Humain**, AGI+, AGI++

Les risques de l'AGI (2)

Scenarios selon Bugaj [3] :

- Faible risque
  - AGI Accompagnée d'une intelligence n'excédant pas un certain niveau pré-determiné;
  - Soft takeoff de l'AGI
- Fort risque
  - Hard takeoff
    - On est pris de court;
    - Pas le temps pour des corrections
    - Pas le temps d'élaborer un certain contrôle

UNIVERSITÉ LAVAL

Que faire ? Les recherches sont en cours

- Propositions pour des actions venant de la société
- Propositions pour des contraintes externes sur le comportement de l'AGI;
- Recommandations pour la conception d'AGI avec des contraintes internes sur leur comportement.

Vers des contraintes venant de la société

- Ne rien faire
- Intégrer les AGI à la société
- « règlementer » la recherche sur les AGI
- Booster les capacités/habiletés humaines
- Renoncer à la technologie qui peut mener à l'AGI.

Ne rien faire

- AGI est trop loin pour que l'on s'y intéresse ;
  - Pas sure
- Peu de risques de l'AGI et donc pas d'action;
  - Les risques existent
- Laissons l'AGI nous détruire : place à plus forts (plus rationnels) que nous !
  - Bien des valeurs humaines d'aujourd'hui sont à préserver.
  - Certaines sont à changer et d'autres à améliorer pour une humanité future « meilleure ».

Intégration de l'AGI à la société

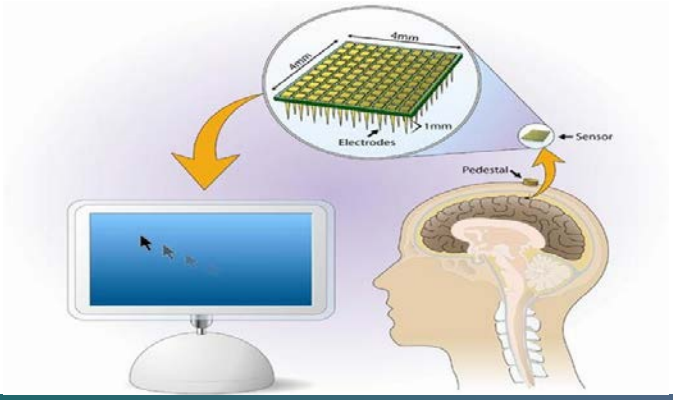
- Contrôler l'AGI d'un point de vue légal et économique ;
  - AGI seront tellement intégrés aux humains que ça risque d'être insuffisant;
- Encourager les « valeurs positives » au niveau des sociétés humaines
  - Les AGI auront tendance à apprendre à se comporter comme des humains

Règlementer la recherche sur l'AGI

- Encourager la recherche sur l'AGI sans danger(sans) (OpenAI) ;
- Passer à une surveillance de masse à l'échelle mondiale (OpenAI) ;
- Encourager le « développement technologique différentiel »
  - (wiki)…is a strategy proposed by transhumanist philosopher Nick Bostrom in which societies would strive to retard the development of harmful technologies and their applications, while accelerating the development of beneficial technologies, especially those that offer protection against the harmful ones.[1]

Booster les capacités/habiletés humaines

- Les humains doivent être (au moins) au même niveau que les AGI
  - **téléchargement de l'esprit** (*Mind Uploading*) est une technique hypothétique qui pourrait permettre de transférer un esprit d'un cerveau à un ordinateur, en l'ayant numérisé au préalable. Un ordinateur pourrait alors reconstituer l'esprit par la simulation de son fonctionnement, sans que l'on ne puisse distinguer un cerveau biologique « réel » d'un cerveau simulé (Wiki).



Booster les capacités/habiletés humaines

- Mind Uploading (Pros et Cons)
  - Copier les cerveaux des meilleurs travailleurs pour avoir un avantage économique;
  - On pourrait copier et mélanger les esprits …
  - Humour, Amour, etc. risquent de disparaître
- Seront nous changés ? Restons nous des humains ?
- Mind Uploading avant ou après l'avènement de l'AGI? Si avant risque t-elle de booster AGI?

UNIVERSITÉ LAVAL

Renoncer à la technologie menant à l'AGI

- L'AGI est-elle du même niveau que certaines manipulations génétiques ?
- Bannir l'AGI, la restreindre ou la déclarer ``hors la loi''.
- Restreindre le commerce et l'utilisation d'un certain harware menant à l'AGI;

UNIVERSITÉ LAVAL

Que faire ? Les recherches sont en cours

- **Propositions pour des actions venant de la société**
- Propositions pour des contraintes externes sur le comportement de l'AGI;
- Recommandations pour la conception d'AGI avec des contraintes internes sur leur comportement.

UNIVERSITÉ LAVAL

Contraintes externes

- Confiner les AGI dans un certain environnement ou les entrées/sorties sont très contrôlées;

- Des AGI qui peuvent détecter les mauvais comportements d'autres AGI
  - Pas évident

Contraintes internes

- AGI uniquement comme Oracle AI
- Top-down safe AGI
- Bottom-up hybrid safe AGI
- Affaiblir la portée des buts à réaliser et trouver le moyen de relâcher aussi la persistance à réaliser des buts.
- Permettre aux AGI de faire des vérifications formelles (conformément à la sureté, sécurité, alignement des objectifs, etc.)

AGI uniquement comme Oracles

- Dans ce cas, l'AGI ne fait que répondre à des questions et il n'exécute pas d'actions.
  - Oracles peuvent devenir des experts: Apache (un Système Expert) donne ses conseils, les médecins lui font de plus en plus confiance, à la fin ils exécutent les yeux fermés ses recommandations.
  - Les utilisateurs sont tentés à émanciper les oracles et à les rendre des décideurs autonomes
- Les oracles pourraient être facilement détournés pour exécuter des actions (par les autres en particulier)

Top Down Safe AGI

- Prendre une théorie éthique spécifique et l'implémenter
  - Les 3 lois robotiques d'Asimov
  - Axiome éthique universel
  - Axiome Utilitarien du genre : Prendre les actions qui vont dans le sens d'un plus grand bien-être des humains tout en diminuant leur « souffrance ».
  - Apprendre les valeurs (via une fonction d'utilité) qui peuvent matcher les préférences des humains.

Bottom-Up Safe AGI

- Ce genre d'approche est généralement basée sur l'évolution et vise a créer une AGI qui évolue en apprenant les valeurs humaines
  - Apprentissage par renforcement
  - Réseaux de neurones

Autres possibles dangers

- Les humains munis de chips AGI
- Multi-agents AGI
- Multi-agents AGI +
  - Imprimante 3D
  - Big Data
  - Cloud
  - Google’s Computer Quantics ou autres
  - Nanotechnologie

AGI comme espoir

- Planète en meilleur état;
- Utilisation efficace des ressources;
- Économie très boostée;
- Prolongation de la durée de vie;
- Éducation at large;
- Recul de la pauvreté;
- Recul de la violence.

UNIVERSITÉ LAVAL

Références

[1] Sotala, K. and Yampolskly, R. V. Responses to Catastrophic AGI Risk: A survey. *Physica Scripta* 50, 2015. Contient beaucoup de références intéressantes.

[2] Bugaj, S. V. and Goertzel B. Five Ethical Imperative and their Implications for Human-AGI Interaction. *Dynamical Psychology (2007)*.

[3] Penrose R. *Les ombres de l'esprit,* Intereditions, 1995.

[4] Bostrom, N. *Superintelligence: Paths, dangers, Strategies*, Oxford, 2014.

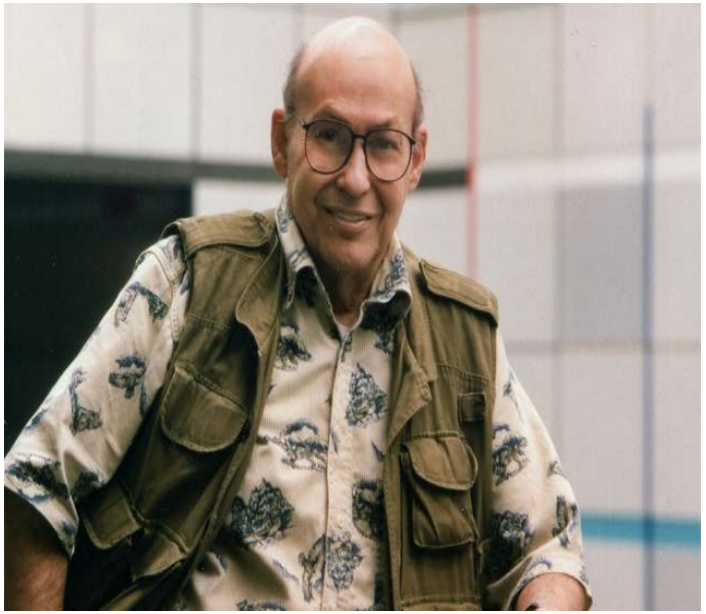
UNIVERSITÉ LAVAL

Dernière nouvelle

- Coup de maître dans l'univers de l'intelligence artificielle. Pour la première fois, un ordinateur a battu un joueur de go professionnel, comme le détaille un article de recherche publié dans la revue Nature du jeudi 28 janvier 2016.

- Un programme de Google DeepMind a relevé le défi du jeu de go, sur lequel planchent depuis des décennies les chercheurs en intelligence artificielle. Une étape historique.

Salut Marvin



Marvin Minsky, “father of artificial intelligence,” dies at 88

Professor emeritus was a co-founder of CSAIL and a founding member of the Media Lab.

MIT Media Lab

January 25, 2016

Adept du Mind Uploading