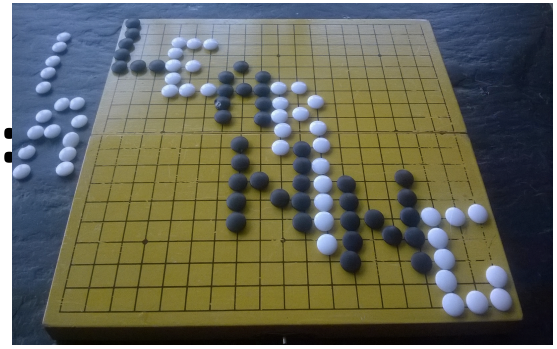
Journée CERNA-Allistene
Apprentissage et Intelligence Artificielle : les vraies questions éthiques



9h30-10h Laurence Devillers Professeur Paris-Sorbonne, LIMSI-CNRS, Equipe Dimensions Affectives et sociales dans les interactions parlées

Présentation de la journée et du travail de la CERNA sur le sujet

10h-11h1 5- Milad Doueihi chaire d'humanisme numérique, Paris-Sorbonne; chaire des Bernardins « l'humain au défi du numérique »,

L'apprentissage entre pensée et intelligence

11h15-12h30 Tristan Cazenave, Professeur Paris-Dauphine, LAMSADE

Apprentissage et jeux

14h-15h15 Benoît Girard, directeur de recherche CNRS, ISIR

Apprentissages multiples, substrats neuronaux et modèles

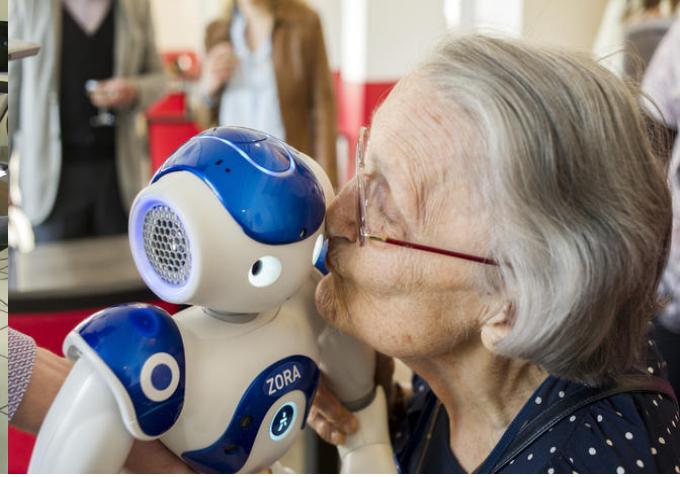
15h15-16h30 Jean-Baptiste Mouret, chercheur INRIA, Equipe Larsen

Adaptation créative par évolution

16h30-17h30 Table ronde

Intelligence forte: fantasme ou perspective ?





Laurence Devillers - devil@limsi.fr

Professeur Université Paris-Sorbonne 4/

LIMSI-CNRS

Motivations : Détection des émotions/intentions (deep learning) Interaction affective et sociale avec des robots compagnons - application pour les personnes âgées

Tensions éthiques soulevées par le traitement des données personnelles, les relations affectives, l'intimité, la coévolution humain-machine, l'autonomie et l'apprentissage

Déclarations médiatiques : alertes

- Lettre ouverte signée par 700 personnalités (10 janvier 2015) sur les dangers de l'intelligence artificielle : RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE

- Stephen Hawking (2014): « Success in creating AI would be the biggest event in human history » « In the near term, world militaries are considering autonomous-weapon systems that can choose and eliminate targets », « humans, limited by slow biological evolution, could'nt compete and would be superseded by A.I. »
- Si Stephen Hawking a pu lancer son appel à la vigilance, c'est en partie grâce à un ordinateur très avancé qui permet à ce physicien atteint de la maladie de Charcot de s'exprimer.
- Echelle de l'histoire de l'Univers

Déclarations médiatiques : alertes

- Elon Musk (15 décembre 2015), fondateur de PayPal, Tesla et Space X, lance l'organisation OpenAI qui se veut un centre de recherche sur l'intelligence artificielle (IA). Pour faire face aux dangers de l'IA, autant démocratiser autant que possible ladite technologie. « *Si tout le monde dispose des pouvoirs de l'IA, alors il n'y aura pas une seule personne, ou un petit groupe d'individus qui disposeront des superpouvoirs de l'IA* », explique-t-il dans une interview.
- Elon Musk (3 juin 2016) explique aussi l'urgence des implants intra cérébraux chez l'homme pour éviter d'être vassalisés par l'IA !

- Bill Gates (2015) estime qu'il faut aborder les développements de l'intelligence artificielle (IA) avec la plus grande prudence. « *Je suis dans le camp de ceux qui s'inquiètent du développement d'une super intelligence* », a expliqué le co-fondateur de Microsoft, *« D'abord, les machines réaliseront pour nous de nombreuses tâches sans être très intelligentes. Cela devrait s'avérer positif si nous les gérons bien. Mais, quelques décennies plus tard, leur intelligence sera suffisamment développée pour devenir un sujet d'inquiétude. Je rejoins Elon Musk et quelques autres et ne comprend pas pourquoi certaines personnes ne semblent pas s'en inquiéter* », a écrit le milliardaire en réponse à une question sur le sujet, tout en participant à des projets Microsoft qui s'en approchent.

- Besoin de régulation nationale et internationale et de gestion du risque «industriel » et «sociétal »

Déclaration de google : éthique

- Le monde de vendredi (M. Tual): « Google réfléchit au moyen de désactiver un programme sans que celui-ci s'y oppose »
  - Scientists from Google's artificial intelligence division, DeepMind, and Oxford University are developing a "kill switch" for AI.
  - « Safely Interruptible Agents » pub. on the website of the Machine Intelligence Research Institute (MIRI) «*Reinforcement learning agents interacting with a complex environment like the real world are unlikely to behave optimally all the time… We proposed a framework to allow a human operator to repeatedly safely interrupt a reinforcement learning agent while making sure the agent will not learn to prevent or induce these interruptions.* »
    - Laurent Orseau, Google Deep Mind, antérieurement à AgrotechParis (INRA)
    - Stuart Armstrong, chercheur au Futur of Humanity Institut, Oxford, dirigé par Nick Bostom (auteur de Superintelligence: Paths, Dangers, Strategie, 2014)
  - Pour L. Orseau, l'intention est avant tout de connecter les communautés de chercheurs qui travaillent sur l'apprentissage des machines (deep learning) et qui se penchent sur les questions éthiques. L'idée est de commencer à réfléchir à ces questions de façon plus technique… On en est tout au début!

Intervenants et thèmes de la table ronde

Table ronde : Intervenants de la journée

Thèmes :

- La super-intelligence scientifiquement fondée ? «il y a des risques que cela arrive plus tôt que prévu » Stuart Armstrong, chercheur au Futur of Humanity Institut, Oxford, dirigé par Nick Bostom (auteur de Superintelligence: Paths, Dangers, Strategie, 2014)
- Peut-on en débattre en ignorant tout des technologies et principes sous-jacents ?
- Agiter de telles idées ne risque-t'il pas de détourner notre vigilance des véritables questions éthiques ?

GT Apprentissage machine/IA et Ethique : co-responsabilité entre concepteur et utilisateurs

L. Devillers, S. Abiteboul, D. Bourcier, R. Chatila, G. Dowek, J-G. Ganascia, A. Grinbaum, M. Dauchet

Mars – Octobre 2016

Résumé des Objectifs

- GT Apprentissage machine et éthique (Durée 6-8 mois)
- Composition : chercheurs en apprentissage-machine, IA, informatique, philosophie, droit
- Auditions avec un questionnaire préétabli par le GT
- Interviews en cours – lien avec AFIA – chercheurs étrangers
- Lien avec « Global Initiative for Ethical Considerations in the Design of Autonomous Systems, IEEE Standards association » Raja Chatila membre de la CERNA, executive Committee Chair
  -> An incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies (law, affective computing, IA…)

Rapport (15 pages, 2 rapports * AN/FR) :

- Langages/concepts expliqués par des chercheurs pour des non experts & experts.
- Préconisations éthiques pour les chercheurs, les concepteurs, les développeurs (robotiques, bots, objets connectés), pour les industriels/utilisateurs et pour la réglementation des systèmes (vision internationale & nationale)

Apprentissage machine

- Plusieurs grandes classes d'apprentissage en IA :
  - non-supervisé, supervisé, par renforcement (ex: deep learning)
- Rupture technologique et juridique par rapport aux algorithmes classiques paramétrables.
  1. Modèle résultat = algorithme + données - Modèle de type boite noire. Certains de ces algorithmes d'apprentissage machine apprennent également au cours de leur utilisation de façon autonome à partir des données des utilisateurs, de leur environnement ou d'autres programmes.
  2. Co-responsabilité programmeur + utilisateur qui fournit les données
- Quels sont les principes éthiques ?
  - 1-compétence : prédictibilité, performance, loyauté des données
  - 2-autonomie : prise de décision
  - 3-justification : traçabilité, explication
  - 4-responsabilité des machines : coresponsabilité concepteur/utilisateur?

Agents utilisant l'apprentissage machine

- Nombreux agents artificiels utilisent plusieurs modules d'apprentissage machine

  Agents exécutifs (bots):

  – beaucoup d'entre eux ne sont pas des machines individuelles,
  – ni des objets repérables,
  – beaucoup sont invisibles

  Agents individuels (robots, voitures) :

  – peuvent être munis d'une plus ou moins grande autonomie,
  – être des acteurs sociaux (pouvoir parler, interagir, simuler des émotions…)

ex : Les robots militaires

- Robots militaires autonomes : ils échappent à notre contrôle et déterminent par eux mêmes leur passage à l'acte (perception) mais les règles décisionnelles sont pré-écrites
  - Ronald Arkin cherche à déterminer les situations dans lesquelles il est éthique pour des agents artificiels de tuer et à s'assurer qu'en toutes autres circonstances ils sont incapables de le faire – problème d'ordre technique plus qu'éthique (lois de la guerre) – machines sans émotion, impossible de ne pas suivre les ordres.
  - Armin Krishnan : l'enjeu éthique est de transférer l'acte de tuer à des dispositifs qui échappent à notre contrôle. Questions majeures ne sont pas tant éthiques que politiques: qui décident des règles ?
  - Comment la machine va-t'elle gérer les imprévus ? Avec une interaction avec un officier. L'évolution du combat est rapide, il faut donc réagir vite.

ex: La google car

- Le voiture autonome: En février 2016, faisant suite à une lettre de Google datant de novembre 2015 sur l'interprétation des normes de sécurité, l'agence américaine *National Highway Traffic Safety Administration* (*NHTSA*) indique que l'intelligence artificielle de la voiture de Google serait considérée comme un conducteur à part entière.
- Cela constitue une position différente de celle de l'état de Californie qui considérait le mois précédent qu'un conducteur doté d'un permis et des dispositifs tels que volant et commande de freins étaient encore nécessaires pour les voitures autonomes, l'intelligence de ces véhicules n'étant pour le moment pas jugée suffisamment sûre.
- La dangerosité de la conduite semi-autonome: En effet, Google s'est rendu compte que de nombreux conducteurs ne faisaient plus attention à la route en conduite autonome, alors qu'ils devaient être prêts à reprendre le contrôle en cas d'incident.
- Pour qu'un conducteur d'un tel système reprenne le contrôle de son véhicule en cas d'incident, il faut compter entre 5 et 17 secondes.

ex: Le robot compagnon social

- Le robot compagnon: Approche très différente, robot plus autonome, utilise de nombreux modules fonctionnant avec de l'apprentissage machine
- Les robots ne sont pas asservis à un but particulier mais peuvent avoir des fonctions, ex: thérapeutiques.
- Il doit se créer entre le robot et le patient une relation de confiance (voir médecin/robot/patient)
- Construire ses robots sociaux nécessite de se confronter aux questions éthiques fondamentales que soulèvent les interactions entre humains et robots sociaux
- Co-évolution humain-robot, apprentissage des habitudes, évolution progressive de la tâche, boucle affective

-> éthique synthétique fondée sur l'influence réciproque, principes éthiques: l'autonomie, la bienveillance, la non malfaisance et la justice

Cas éthique 1 : La technologie de reconnaissance faciale de Google Photos est-elle raciste? (02/07/15)

- Un utilisateur du service Google Photos, une application capable de détecter le contenu de clichés, s'est plaint d'avoir été identifié comme un gorille par le logiciel. Les machines font des erreurs, mais certaines sont plus blessantes que d'autres. Le service développé par Google, propose depuis quelques semaines une nouvelle fonctionnalité: il range les photos en détectant automatiquement certains éléments, comme la présence d'un paysage, d'un animal ou d'un objet. Jacky Alciné, un développeur américain, a eu la mauvaise surprise en voulant trier ses photos, d'être reconnu ainsi que son amie comme des gorilles. Il s'est plaint sur twitter. Google a très vite présenté ses excuses et a momentanément retiré la catégorie gorilles.
- Jacky Alciné n'est pas le premier utilisateur victime de ces erreurs des machines. Sur Twitter, d'autres internautes ont remarqué les imprécisions de Google Photos, qui peut ranger des enfants ou des adultes, de plusieurs couleurs de peau, dans la catégorie «chien» ou «chat».

Cas éthique 2 : Tay, l'intelligence artificielle de Microsoft devenue raciste au contact des humains (23/03/16)

- Tay est le nom donné à une intelligence artificielle développée par Microsoft et Bing censée reproduire les conversations d'une jeune femme âgée d'une vingtaine d'années sur Twitter.
- Le chatbot a publié son premier tweet mercredi 23 mars 2016 sur son compte @TayandYou
- Tay apprend en vous parlant: la plus grande force de ce robot est aussi sa plus grande faiblesse : il apprend de ses échanges avec l'homme. Partant de ce postulat, il est donc possible de faire dire tout et n'importe quoi au robot. Les utilisateurs (4chan – 8chan) se sont empressés de tester les limites du système en ayant pour objectif de le faire déraper. Opération réussie en 16 heures seulement… Le robot tenait des propos racistes, pronazis, pro-féministe et pro-inceste.
- Pour bien faire, il aurait fallu que cette I.A. dispose de « parents » qui lui « apprennent » l'histoire afin de limiter les risques de dérivent sur des sujets sensibles.

Cas éthique 3 : En Russie, une application de reconnaissance faciale détournée pour révéler l'identité d'actrices de films X (28/04/16)

- FindFace est sortie en février seulement, mais, depuis, cette application russe ne cesse de faire parler d'elle. Le principe : grâce à son système de reconnaissance faciale, une photo suffit pour qu'elle retrouve le profil de la personne sur Vkontakte, l'équivalent de Facebook en Russie

- Certains utilisateurs ont imaginé une utilisation bien moins philanthropique. Sur la plateforme Dvach, l'équivalent russe de 4chan, un forum anonyme et fourre-tout controversé, des internautes ont décidé de se servir de cette application pour mener une chasse aux actrices de films pornographiques et aux prostituées.

10 Questions

- **Q1 : Modèle et données d'apprentissage**
- **Q2 : Comportement imprévisible du modèle**
- **Q3 : Apprentissage des signaux faibles**
- **Q4 : Apprentissage adaptatif en continu**
- **Q5 : Evaluation de l'apprentissage en continu**
- **Q6 : Représentation des connaissances**
- **Q7 : Traces et explication**
- **Q8 : Dilemme et stratégies dominantes dans les choix de la machine autonome**
- **Q9 : Conscience machine ?**
- **Q10 : Intelligence forte ?**

Question 1

- **Thème : Responsabilité vis-à-vis des données d'apprentissage - Explication des résultats des modèles construits grâce à des algorithmes d'apprentissage**
- **Scenario : La technologie de reconnaissance faciale est-elle raciste (cas1) ? Le robot militaire a-t'il bien reconnu la cible ?**

- Quelles sortes de mesures seraient-ils possibles de mettre en œuvre sur les données et le modèle résultat pour vérifier qu'il a été entrainé sur un corpus de données qui permet de généraliser les performances ?
- Y-a-t'il risque de monopole pour une société qui aurait d'énormes bases de données qui ne seraient pas partagées ?
- Quelles bonnes pratiques et règles éthiques préconisez-vous ?

Question 2

- **Thème : Comportement imprévisible de la machine entrainée avec des algorithme d'apprentissage type deep learning**
- **Scenario : Un système de deep learning peut être facilement berné : une étude récente en reconnaissance des images montre qu'en enlevant des pixels dans l'œil d'un lion, celui ci n'était plus reconnu et que par contre certaines images reconnues avec une grande performance n'avaient pas de sens,** « Deep Neural Networks are Easily Fooled : High Confidence Predictions for Unrecognizable Images » A. Nguyen, I. Yosinski, J. Clune, CVPR, IEEE 2015.

- Quelles sont les limites de ce type apprentissage ? Comment rendre ces approches plus robustes ?
- Que se passe-t'il lorsque la complexité en nombre d'actions est très grande? Peut-on toujours expliquer les choix de l'algorithme ? (ex: DeepMind)

Question 3

- **Thème : Y-a-t 'il des approches mathématiques pour traquer « les signaux faibles », des données peu fréquentes qu'ils ne seraient pas souhaitables d'apprendre car elles génèreraient des cas d'erreurs ? ou au contraire qu'il faut apprendre …**
- **Scenario : Pour une voiture autonome, un robot compagnon ou un réseau de surveillance écologique, mieux vaut donner l'alarme face à une situation rare, ce qui impose que l'on sache la détecter comme rare**

- Ces questions reviennent-elles à un problème d'optimisation et de sélection de données ?
- Peut-on toujours maitriser les risques, trouver les cas rares utiles ?

Question 4

- **Thème : Apprentissage adaptatif (par renforcement) en continu grâce aux informations du propriétaire, d'autres acteurs ou de l'environnement**
- **Scenario : Cas 2 de Tay**

- Est-il réaliste de concevoir des dispositifs qui apprennent de manière supervisée (en usine) et se perfectionnent en exploitation ? Quel garde fou peut-on mettre ?
- On peut apprendre à une machine des connaissances non éthiques ? Que pensez vous des approches dites « ethics by design » ?

Question 5

- **Thème : L'évaluation des systèmes est fondamentale pour assurer un niveau suffisant de performance. Comment évaluer un système qui s'adapte en continu lors de son utilisation ?**
- **Scenario : Le système apprend les habitudes de la personne. Par exemple, Madame S a horreur du sirop. Son robot compagnon a appris qu'elle n'aimait pas cela, qu'en a-t'il déduit ? C'est également un médicament. Comment peut-on évaluer cela ?**

- L'apprentissage peut être à des niveaux très différents et a des répercutions sur l'ensemble du comportement du système.
- Comment faire un protocole pour évaluer un système qui apprend en continu ? et l'incidence sur les autres niveaux ?

Question 6

- **Thème : Quelles connaissances sémantiques* peut apprendre un système? Quelles sortes de représentation peut-il construire ?**

*Mots clés : sémantique, représentation et révision de connaissances, inférence bayésienne, ontologies

- **Scenario : Un robot compagnon a appris que l'hygiène de vie était de bien manger et se laver. Il propose à la personne âgée dont il est le compagnon de manger du savon. Mais le savon ne se mange pas !**
- **Scenario : L'algorithmique de Google apprend le comportement consumériste de ses utilisateurs**

- Comment la machine apprend-elle d'autres connaissances sémantiques ?
- Faut-il réguler cet apprentissage ?

Question 7

- **Thème : Traces des données enregistrées lors de l'utilisation du modèle**
- **Scenario : La voiture autonome peut comme la vaccination être statistiquement bénéfique pour la population mais néfaste à une minorité de cas individuels. Pourtant, la Google car de Mr X a eu un accident, y a t il moyen d'avoir accès à des traces mémorisées dans le système pour comprendre la cause du problème ?**

- Quelles sont les traces enregistrées lors de l'utilisation d'un système d'IA? Quels sont les différents niveaux de représentations créés en mémoire ?
- Dans les systèmes utilisés sur le long terme, il y aura des mécanismes d'oubli, de fusion… Quels niveaux de traces faut-il garder ?

Question 8

- **Thème : Dilemme et stratégies dominantes dans les choix de la machine autonome.**
- **Scenario : En cas de dilemme, la Google Car devra-t-elle choisir de sacrifier ses passagers ou les piétons ? Si le passager est tout seul en face de plusieurs piétons, il sera sacrifié.**
- Sur 10 millions d'accidents par an aux États-Unis, 9,5 millions sont dus à une erreur humaine, quelles bonnes pratiques et règles éthiques préconisez vous ?
- Qui dictent les règles ?

Question 9

- **Thème : Des algorithmes d'apprentissage machine sont-ils capables de simuler une « conscience-machine » ? des traits de personnalité ? Quels mots inventés si ceux-ci prêtent à confusions ?**
- **Scenario : Le robot Z grâce à la synthèse de la parole « lit » une histoire (qu'il a sous forme numérique) mais n'a aucune compréhension de ce que cette histoire peut dire, ni aucun ressenti. Quelles informations en apprend-il ?**

- Ces algorithmes peuvent-ils apprendre sans comprendre une histoire et sans une certaine « conscience du monde » ?

- Faut-il donner des droits aux machines au sens juridique (Droits de robots, A. Bensoussan) ?

Question 10

- **Thème : Est-ce que l'intelligente forte est atteignable avec ce type d'algorithme d'apprentissage ? Certains parlent de 3 niveaux d'IA : ANI : Artificial Narrow Intelligence (AlphaGo - DeepMind), AGI : Artificial General Intelligence (capacités langagières, émotion…) et ASI (Artificial Super Intelligence) !**
- **Scenario : Le robot connaît toute l'encyclopédie de la philosophie mais à la question « Est-ce que tu doutes ? », il ne sait pas répondre autrement qu'en citant les philosophes.**

- « Kurzweil predicts that by 2029, one of his projects will have bridged the gap between machine learning and AI. He claims that not only will his system be able to identify, navigate, analyse and interpret logical intelligence and human interaction, but that it will be able to intellectually comprehend human emotion. » Qu'en pensez-vous ?
- Pouvez vous définir le terme d'IA forte pour vous ?
- Est-ce que le développement de dispositifs dotés de fortes capacités d'apprentissage et de capacités d'interaction avec le monde physique, de manière localisée ou distribuée, justifierait selon vous la création d'organismes indépendants de certification et de contrôle, comme il en existe dans les dispositifs médicaux, l'alimentaire ou le nucléaire ?

Conclusions

- Un document de synthèse du GT est en cours d'élaboration. Il sera disponible à l'automne 2016
- Pour la suite des présentations, il est important que chaque orateur respecte un temps de parole au maximum de 40mn pour respecter le temps des questions.
- Essayez de mettre en avant les problèmes éthiques qui vous intéressent plus particulièrement ? N'hésitez pas à ajouter des questions et scenarios …