TESaCo

Les Cahiers de TESaCo N°5

Février 2024

TESaCo

Les Cahiers de TESaCo N°5

IA ET ROBOTIQUE /
SOUTENABILITÉ DE L'IA /
APPRENTISSAGE PROFOND /
SAGESSE COLLECTIVE /
TRAITEMENT DES ÉMOTIONS

Technologies émergentes et sagesse collective

Comprendre, faire comprendre, maîtriser

Février 2024

ISBN : 978-2-9518704-5-1

INSTITUT DE FRANCE 1795 1832

ACADÉMIE DES SCIENCES MORALES ET POLITIQUES
INSTITUT DE FRANCE

fondation Simone et Cino Del Duca INSTITUT DE FRANCE

CONTACT

Vous pouvez nous contacter à l'adresse courriel suivante tesaco@asmp.fr ou via notre site internet https://www.tesaco.fr/

Les Cahiers de TESaCo N°5

Directeur de publication : Daniel Andler
Achevé d'imprimer par Corlet, 14110 Condé-en-Normandie
Dépôt légal : octobre 2025 — *Imprimé en France*
ISBN : 978-2-9518704-5-1

TESACO

En l'espace de deux décennies, les technologies dites émergentes — biotechnologies, technologies de l'information et de la communication, technologies issues des neurosciences cognitives, nanotechnologies…— ont profondément modifié les conditions d'existence à l'échelle planétaire et affecté tous les secteurs d'activité humaine. Porteuses de solutions mais aussi de menaces pour nos équilibres fondamentaux, ces nouvelles technologies sont devenues si puissantes qu'on ne sait comment en reprendre le contrôle, alors même qu'elles continuent de se développer, ouvrant la voie à des conséquences et à des risques imprévisibles.

Cet état de fait appelle un effort pour mieux comprendre les technologies et leurs effets, informer le public et les responsables politiques, et proposer des dispositifs pouvant contribuer à maîtriser l'évolution en cours.

L'Académie des sciences morales et politiques a souhaité participer à cet effort, et avec l'appui de la Fondation Simone et Cino del Duca elle a lancé en 2019 le cycle d'études « Technologies émergentes et sagesse collective » (TESaCo).

LES CAHIERS DE TESACO

Les Cahiers de TESaCo est une publication périodique qui présente les travaux de l'équipe du projet, organisée en six groupes de travail thématiques : biotechnologies, intelligence artificielle et robotique, sciences cognitives appliquées, libertés-éthique-droit, numérisphère, anthropologie numérique.

LE COMITÉ ÉDITORIAL

SOMMAIRE

INTRODUCTION

À PROPOS DU CAHIER N°5

SERENA CIRANNA

Serena Ciranna est chercheuse post-doctorale en sociologie du numérique au Southern Center for Digital Transformation de l'Université Federico II de Naples. Ses recherches portent sur la mémoire numérique et sur le rôle des systèmes algorithmiques dans la construction d'identités narratives. Elle a été pendant cinq années assistante de recherche pour le projet TESaCo.

Depuis ses débuts, le cycle d'études TESaCo - Technologies émergentes et sagesse collective - publie une partie de ses travaux dans ses *Cahiers*. Ils sont une trace de ce projet qui a commencé en 2019 et qui verra sa conclusion en 2024. Des rencontres, des séminaires, des auditions et des entretiens sont à l'origine des textes publiés dans les *Cahiers de TESaCO*. Dans ces documents, le lecteur pourra trouver des pistes de réflexion émergeant d'un dialogue qui se poursuit depuis presque quatre ans entre des chercheurs de domaines différents qui unissent leurs efforts pour élaborer une forme de sagesse collective apte à évaluer l'impact des technologies émergentes sur la société et à frayer des chemins conduisant de la réflexion aux actions et aux transformations.

Le présent numéro 5 des *Cahiers* est entièrement consacré au sujet de l'IA, technologie émergente par excellence – si l'on peut dire –, non seulement du fait de la rapide expansion de ses applications durant les toutes dernières années, mais aussi en raison de la complexité de sa progression historique – par des bonds » et par des approches différentes qui ont abouti aux impressionnantes perspectives actuelles –, et de par la multitude de scénarios que les intelligences artificielles (avec la difficulté de maintenir ce terme au singulier) laissent entrevoir.

La question des possibles impacts positifs et négatifs des intelligences artificielles sur la société occupe une place centrale dans la réflexion menée au sein de TESaCo. Cet intérêt s'est concrétisé dans l'organisation de nombreux événements scientifiques, dont

des séminaires internes, des colloques en France et à l'international et des publications, en particulier l'ouvrage de Daniel Andler, *Intelligence artificielle, intelligence humaine: la double énigme* paru en mai 2023 et qui a rencontré un écho important dans la presse. Dans ce cadre, un effort considérable a été déployé dans la réalisation d'auditions d'éminents spécialistes du domaine. Trois des quatre textes présents dans le *Cahier* n°5 sont issus de ces auditions, dont certaines ont été parallèlement publiées sous forme de vidéo sur le site web de TESaCo. Il s'agit des auditions de Raja Chatila, Michèle Sebag et Yann LeCun. Le travail de Mehdi Khamassi, qui a réalisé les auditions, et des membres de TESaCo qui y ont participé, a permis de mettre en place un dialogue serré, bien que virtuel, entre les auteurs de ces textes et leurs différents angles d'approches.

Ces trois auditions sont complétées par le texte issu du séminaire interne animé par Célia Zolynski sur la régulation du traitement des données émotionnelles issues de la reconnaissance automatique des émotions, un ensemble de techniques qui présentent des risques bien concrets pour les individus. Les propositions contenues dans le texte nous font d'ailleurs entrevoir le nouveau domaine de la défense des « droits cognitifs » des individus, pour les protéger des risques de manipulation liés à des technologies de plus en plus invasives.

Ces quatre témoignages d'acteurs centraux de la recherche et de la régulation de l'IA font émerger un portrait détaillé des promesses et limites, ainsi que des questions éthiques et légales posées par les tout derniers développements qu'on appellera, dans la terminologie recommandée par Daniel Andler (2023), Raja Chatila et d'autres auteurs des « systèmes artificiels intelligents ».

Chaque texte de ces *Cahiers* n°5 touche à certaines questions spécifiques dans le domaine de l'IA : la robotique et la cognition incarnée (audition de Raja Chatila), la soutenabilité de l'IA (audition de Michèle Sebag), les évolutions de l'apprentissage profond et les potentiels positifs de l'IA (Yann LeCun), la reconnaissance automatique des émotions (séminaire de Célia Zolynski). Dans leur ensemble, les quatre textes répondent de manière transversale à des questions qui touchent à trois catégories de problèmes liés à l'IA : la recherche, l'impact social et la politique.

Le sujet de la recherche en IA, abordé dans les trois auditions, porte sur des aspects inhérents au développement des systèmes artificiels intelligents, dont notamment le dialogue interdisciplinaire qu'il nécessite et suscite. Ce faisant, les auditions ont toutes souligné le dialogue de la recherche en IA/robotique avec les neurosciences et sciences cognitives et recherche en IA et robotique, les limites et possibilités de développement ultérieures des modèles actuels, les promesses et illusions d'une IA « forte », le problème de la sémantique des agents conversationnels, l'autonomie des machines, pour en citer seulement quelques-uns. Les textes des auditions ne manquent pas d'ailleurs d'évoquer les étapes importantes de l'histoire de l'IA, dont les vicissitudes servent dans certains cas de leçon pour concevoir des pistes de recherche futures.

Des préoccupations croissantes concernent l'impact des systèmes artificiels intelligents sur l'humain, entre autres en termes d'influence sur la sphère épistémique et psychologique (désinformation, manipulation) ainsi que pratique (travail, surveillance) des individus. Les nombreuses questions traitées dans ces quatre textes visent alors à évaluer l'impact de l'IA sur la société au sens large, dans le court et long terme – une question générale et « difficile » à laquelle les auteurs ont pu répondre en mobilisant leurs différents champs d'expertise et en nous éclairant sur les risques illusoires ou concrets représentés par l'IA, ceux-ci allant du supposé danger d'extinction de l'espèce humaine au risque de diffusion de la désinformation, à la fracturation de la société due à la disparition d'une vérité partagée ou encore à la menace à la démocratie.

Liée à l'impact social de l'IA, mais distincte de celle-ci, est la question de la politique, cette

dernière devant faire face aux pressions économiques et aux rapports entre les États dans le développement et la régulation de l'IA, ainsi qu'à la tension entre la sphère publique et les entreprises privées du numérique. Des réflexions spécifiques concernent l'effort de régulation et son efficacité, le rôle des comités d'éthique à travers le monde, le rapport entre experts, citoyens et décideurs. Alors, comment mobiliser une sagesse collective pour limiter les risques d'un côté et pour mettre réellement l'IA au service de la société de l'autre ? Comment impliquer les citoyens dans ce processus ? Un élément évoqué au fil des différents témoignages est celui de l'éducation, notamment de l'importance de la formation tant du large public à l'IA que des chercheurs et développeurs d'IA sur les questions éthiques posées par celle-ci.

Alors que le débat sur l'IA ne cesse pas de se renouveler, les textes publiés dans ce Cahier n°5 permettent de s'arrêter sur des éléments importants du discours actuel, de remettre en question des idées reçues et de poser des questions éminemment politiques. Il s'agit, en même temps, d'une plongée dans des problèmes spécifiques de la recherche sur les systèmes artificiels intelligents et d'une exploration plus large sur les questions que l'IA pose à nos sociétés. Ce mouvement de mise en perspective et de prise de distance critique à partir du débat scientifique représente le propre de la contribution que TESaCo souhaite apporter à cet important débat de notre époque.

Les mêmes questions d'éthiques se posent-elles pour l'IA et la Robotique ?

Audition de Raja Chatila

RAJA CHATILA

Raja Chatila est professeur émérite de robotique, d'intelligence artificielle et d'éthique à Sorbonne Université. Auparavant, il a été directeur de recherche au CNRS et a dirigé deux grands laboratoires du CNRS : le Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS) à Toulouse, et l'Institut des Systèmes Intelligents et de Robotique (ISIR) à Paris. Ses domaines de recherche portent principalement sur la robotique autonome, et la robotique cognitive et interactive. Président de la société savante IEEE Robotics and Automation Society de 2014 à 2015, il préside une initiative internationale IEEE pour l'éthique dans l'intelligence artificielle et les systèmes autonomes. Il a participé à plusieurs travaux au niveau national et européen sur les implications éthiques et sociétales de la robotique et est membre du Comité National Pilote d'éthique du Numérique.

L'audition a été menée par Mehdi Khamassi.

1. Les mêmes questions d'éthiques se posent-elles pour l'IA et la Robotique ?

Mehdi Khamassi [0.16]

Bonjour Raja Chatila. Merci beaucoup d'avoir accepté cette audition pour le projet TESaCo, Technologies émergentes et sagesse collective, de l'Académie des sciences morales et politiques, dirigé par Daniel Andler, qui est ici présent. Il y a d'autres membres de l'équipe de TESaCo, Anne Le Goff, Margaux Berrettoni, et moi-même, qui sont là pour te poser un certain nombre de questions.

Juste une brève introduction : tu es professeur émérite en robotique, intelligence artificielle et en éthique à Sorbonne Université, après avoir été directeur de recherche au CNRS. Et tu as été directeur de deux grands labos en France, le LAAS à Toulouse et ensuite l'ISIR à Paris. Tu es très impliqué depuis de nombreuses années sur les questions d'éthique de la robotique, de l'intelligence artificielle notamment, tu étais membre de la CERNA [et actuellement du CNPEN], tu es [président] de l'IEEE

Global Initiatives on Ethics of Autonomous and Intelligent Systems. Et puis plus récemment [en 2018], tu as été nommé par la Commission européenne comme membre du groupe d'experts [qui a fait des recommandations sur des orientations éthiques en vue d'] une proposition de régulation sur l'IA, qui vient d'être publiée [en 2021] afin qu'il y ait consultation [et qui est en cours de discussion au niveau des instances européennes].

Donc c'est au titre de toutes ces réflexions que tu mènes que nous avions envie de te poser des questions, pour enrichir notre réflexion sur les technologies émergentes, leur évolution, et leur impact possible sur la société.

Raja Chatila [1.41]

Bonjour. D'abord merci de m'avoir invité pour cet échange. Je suis extrêmement heureux de pouvoir partager quelques idées avec vous. J'essaierai de vous répondre dans la mesure de mon possible sur ces sujets.

Mehdi Khamassi [2.04]

Une première question qu'on a envie de te poser justement dans le cadre des questions d'éthique qui peuvent se poser en IA ou en robotique. Est-ce que pour toi il y a des particularités ou des points communs à ces questions d'éthique qui font qu'il est important de mettre en avant ?

Raja Chatila [2.23]

Alors, on va être obligé d'aborder la question de la définition. Malheureusement. Aujourd'hui le choix a été fait de manière relativement officielle, d'une certaine façon, d'inclure la robotique dans l'intelligence artificielle. Je dis de manière officielle parce que, au niveau de la Commission européenne en particulier, pour élaborer le plan d'action en intelligence artificielle, il y a eu un choix de définition qui effectivement met dans la même catégorie les systèmes d'intelligence artificielle et la robotique. Bien sûr, est considérée ici une robotique intelligente, au sens où il ne s'agit pas simplement de machines qui sont guidées par des programmes répétitifs ou par des systèmes purement mécaniques. Je pense que c'est une volonté simplificatrice pour ne pas alourdir le langage. Mais après quand on va dans certains détails, on voit certaines différences. De nouveau en termes de politique. Donc je me permets de m'éloigner un peu de l'éthique pour parler de politique, parce que c'est lié.

Dans l'esprit de la Commission européenne en particulier, mais d'autres aussi, la robotique est très liée à l'industrie manufacturière. Cette compréhension de la robotique comme étant des machines qui sont dans l'industrie manufacturière, évidemment, limite grandement l'intelligence des robots. Parce que dans l'industrie manufacturière, on a des tâches répétitives, on a des tâches où on ne demande pas une grande variété, une compréhension vraiment de l'environnement, enfin tous les aspects sémantiques qui peuvent nous intéresser. Mais c'est donc une simplification de dire que la robotique et l'intelligence artificielle, c'est la même chose.

Cette démarche a aussi été adoptée par l'OCDE. La définition de l'intelligence artificielle par l'OCDE aujourd'hui est en train d'être de plus en plus universellement acceptée, à cause de sa simplicité. Elle a été reprise d'ailleurs par la Commission européenne, avec une légère modification, dans la définition qu'elle utilise pour sa proposition de règlement qu'elle a publié le 21 avril [2021], abandonnant ainsi une définition beaucoup plus détaillée, beaucoup plus large, en plusieurs pages, qui a été élaborée par le groupe d'experts que la Commission avait nommé pour l'aider à

élaborer son instrument législatif. Il est intéressant de voir aussi que la définition de l'intelligence artificielle telle qu'elle est donnée par l'OCDE est une définition qui inclut aussi la robotique, puisque cette définition-là parle de systèmes qui peuvent agir y compris sur leur environnement physique. Dans ce sens, ça inclut la robotique, par opposition à un environnement virtuel que de purs programmes informatiques pourraient utiliser. Donc là, il n'y a pas de différence dans cette définition.

Aux origines de l'IA et de la robotique

Raja Chatila [6.31]

D'ailleurs, d'une certaine façon, on retrouve l'absence de distinction entre robotique et intelligence artificielle dans la définition d'origine de l'intelligence artificielle. Parce que le programme de recherche fondateur de l'intelligence artificielle de Dartmouth College ne parlait pas de robotique et d'intelligence artificielle. Il parlait de machines, et de fonctions ou de capacités intelligentes : l'apprentissage, le langage, la formation de concepts, la perception. Ils négligeaient complètement l'action, en revanche. Cependant, dans leur esprit, dans leur texte, il n'y avait pas de distinction entre une machine physique et une machine qui utiliserait uniquement des logiciels. Cependant, ce fait que l'action n'était pas considérée comme vraiment importante dans la manière dont l'intelligence artificielle a été définie et ensuite développée a constitué en quelque sorte un péché originel du programme de recherche en IA. En effet, les travaux se sont focalisés évidemment à ce moment-là davantage sur ce qu'on considérait comme intelligent, sans définir l'intelligence. Donc plutôt des tâches abstraites, et non pas des actions dans le monde réel, même si la perception faisait partie très largement des travaux qui ont été menés dès le départ.

Il faut passer de 1956 à la fin des années 60, avec d'une part Hans Moravec à Stanford – où on voit le premier robot au sens d'une machine physique pilotée par la vision ; son travail essentiel était sur la stéréovision, pour qu'un robot se déplace en évitant des obstacles découverts par la stéréovision ; c'était juste un lien entre la perception visuelle et l'action de déplacement –, pour voir des projets, des programmes, financés par la DARPA, qui s'appelaient Hand-Eye Program, où on essayait de faire une action par un robot manipulateur basé sur la vision, mais surtout avec le projet Shakey au SRI, fin des années 60, début des années 70, avec une publication en 1969 à la première conférence internationale d'intelligence artificielle, qui parlait de ce projet, de ce robot Shakey. Ce dernier n'était pas considéré à proprement parler comme un robot puisque l'article parlait d'un automate : « application des techniques d'intelligence artificielle sur un automate ». Donc on ne parlait pas de robot, et Shakey était, je dirais, le parangon du robot mobile qui découvre son environnement, qui s'y déplace, qui prend des décisions, puisque les premiers systèmes de planification STRIPS, symboliques, ont été élaborés à ce moment-là, et l'algorithme A* a été élaboré dans le cadre du projet Shakey. Mais on était en train de mettre en œuvre des systèmes d'intelligence artificielle. Pour eux, il n'y avait pas de différence. Le robot ne représentait pas une singularité dans le paysage de l'IA.

Il faut remarquer qu'en même temps, au début des années 60, le robot fait son apparition dans l'industrie. UNIMATE, le système qui a été inventé par Engel Burger et d'autres pour faire des actions dans les usines, peinture, soudures, etc., qui était un automate, à purement parler un automate, et pourtant ne s'appelait pas un automate. Il s'appelait un « mover » au début. Et le mot robot a été

utilisé pour des raisons commerciales, car il est plus attractif. C'est comme ça que le terme robot a migré d'Asimov, de la science-fiction, à la science et à la technologie. Mais au départ on considérait des machines et de l'IA. Et ensuite le mot robot a eu des usages beaucoup plus développés dans le domaine dit de la Robotique, qui au départ incluait des chercheurs en IA travaillant sur la robotique, et qui petit à petit se sont rendu compte qu'il y avait un problème pour agir dans le monde réel. Ils se sont rendu compte que les techniques d'IA utilisées jusque-là ne permettaient pas d'appréhender correctement l'interaction et la dynamique avec le monde réel.

C'est là qu'il y a eu aussi, au début des années 80, avec Rodney Brooks en particulier, qui a bien formulé ce problème dans l'un de ses papiers : « Elephants don't play chess », qui est revenu en arrière, disons, sur une critique de l'intelligence artificielle, et en particulier de l'intelligence artificielle symbolique, non pas à l'avantage des approches connexionnistes, mais plutôt avec une vision behavioriste du comportement, et donc de l'intelligence, avec l'idée que des couches comportementales pourraient effectivement mieux modéliser, mieux exprimer l'intelligence d'une machine qui interagit avec son environnement.

Retour à la question de l'éthique

Raja Chatila [13.19]

Pour revenir à la question qui portait sur l'éthique en réalité, et non pas sur la définition, mais j'ai été obligé de passer par la définition, à cause maintenant de l'unification, disons, du concept de machine intelligente, qui inclut robot et systèmes d'intelligence artificielle, on peut dire qu'il n'y a pas de différence fondamentale quand on parle d'éthique. Les problématiques éthiques qui se posent ne sont pas exactement les mêmes. En effet, les robots ne vont pas traiter des masses de données, et ne vont donc pas fréquemment être confrontés aux problèmes de biais que l'on rencontre quand on focalise sur les systèmes d'apprentissage modernes. Mais oui, en même temps, si un robot est une machine qui est un agent conversationnel qui bouge et qui collecte des données, alors on va rencontrer ces problèmes-là.

Donc je dirais qu'en réalité il n'y a pas de différence fondamentale du point de vue éthique, entre les systèmes d'IA pure, disons, agissant dans un environnement virtuel, et de robotique agissant dans un environnement réel.

En revanche, il peut y avoir une différence dans la mise en œuvre, dans l'action, puisque le robot agit. Mais il n'agit que comme conséquence des décisions, d'un raisonnement, d'une élaboration de plans, qui eux sont du domaine de l'IA en général. Le robot agit à partir d'une interprétation du monde à travers sa perception, son interprétation des situations, et c'est là peut-être qu'il y a une différence avec les systèmes virtuels. Je ne veux néanmoins pas limiter l'IA aux systèmes virtuels. Mais en tout cas, les systèmes virtuels ne sont ni situés dans un monde réel qui évolue, ni matérialisés, c'est-à-dire soumis aux contraintes de la physique ou de la mécanique, qu'ils ne connaissent pas. À ce moment-là, leur perception du monde, la perception du monde, leur compréhension, leur interprétation peuvent s'en trouver différente. Est-ce que ça différencie la délibération éthique ? Peut-être dans la mesure où certaines décisions ne pourront pas être effectuées, ne pourront pas être mises en œuvre, par exemple. Ou bien parce que justement la problématique éthique de l'interprétation de situations exige des raisonnements qu'on ne maîtrise pas aujourd'hui. Mais ça reste pour moi de l'intelligence artificielle.

Liens avec la cognition incarnée en sciences cognitives

Mehdi Khamassi [16.40]

Daniel, tu voulais rebondir là-dessus ?

Daniel Andler [16.46]

Oui, je suis très heureux de la réponse de Raja. Je suis d'accord. La seule chose c'est qu'il y a, comme Raja le sait très bien, plutôt du côté des sciences cognitives que de l'IA, cette idée que les sciences cognitives classiques, qui sont quand même pas mal associées au paradigme symbolique en IA, ont complètement loupé le côté incarné. C'est-à-dire qu'on accuse les sciences cognitives du début d'être purement intellectualistes, imaginant un cerveau dans une cuve qui reçoit des informations de la tour de contrôle, et qui transmet des informations aux organes moteurs, et que d'une certaine façon, on passe à côté de quelque chose d'essentiel, à savoir que la cognition est fondamentalement incarnée. Je me demandais comment tu voyais ça : est-ce qu'on peut dire que la robotique est à l'IA disons purement virtuelle, ou purement orientée traitement de l'information, ce que la cognition incarnée serait à la cognition intellectualiste des débuts ? Est-ce qu'il y a un parallèle intéressant, ou tout ça, c'est simplement une mauvaise façon de découper les choses ?

Raja Chatila [18.13]

D'une certaine façon, oui. Mais je vais peut-être insister sur un point. Je pense que l'interprétation du monde réel est impossible pour une IA qui n'est pas incarnée. Je pense que comprendre le monde est inhérent au fait qu'on puisse y agir. Donc ce n'est au fond pas du tout la même forme d'intelligence. C'est de l'intelligence artificielle si on veut, quand on parle de systèmes, de machines. Mais ce n'est pas du tout la même forme d'intelligence. Des intelligences artificielles – je n'aime pas du tout mettre des articles comme ça, en disant « une intelligence artificielle » –, enfin, un système d'intelligence artificielle qui ne fait que traiter des données pour faire des classifications ou pour faire n'importe quoi, ne pourra pas appréhender la signification du monde physique. Seule la matérialisation pourrait éventuellement le permettre. Et ça, ça crée une différence importante : finalement la « véritable intelligence » au sens propre, au sens de compréhension, ne peut être qu'incarnée.

Non pas une IA mais des IA

Daniel Andler [20.04]

OK, merci. Très bien. Je suis très heureux de voir que tu n'aimes pas « une IA », mais « des IA ». Parce que moi-même je déteste ça, et j'ai inventé mon propre sigle. Mais on est tout à fait d'accord. J'appelle ça des « SAIs », des systèmes artificiels intelligents.

Raja Chatila [20.22]

Tout à fait. Je dis aussi « systèmes d'intelligence artificielle ». C'est ce que j'utilise le plus souvent, et là je te rejoins complètement.

Daniel Andler [20.35]

OK, parfait.

Interprétation du monde par une IA forte

Mehdi Khamassi [20.38]

Pour rebondir sur ça, sur la question de l'interprétation du monde que pourrait faire un agent artificiel, et si possible un agent incarné, donc de compréhension et de raisonnement sur le monde, on arrive vite à des questions de potentiel développement de ce que certains appellent une IA forte. Est-ce que c'est quelque chose que tu considères comme possible ? Est-ce que tu es sceptique sur ça ? Est-ce que tu trouves que ce terme convient ou pas ? Qu'est-ce que tu en penses ?

Raja Chatila [21.26]

Je n'aime pas cette classification d'IA forte et IA faible, etc. Assez paradoxalement, c'est drôle, car à l'origine la distinction a été faite, je crois, pour justement prouver que l'IA forte n'existe pas. C'est John Searle qui désignait l'IA forte comme une IA qui est capable de faire des raisonnements comme humain, alors que l'IA faible n'en faisait pas, pour dire qu'il n'y avait que de l'IA faible en réalité. Je résume. Et maintenant c'est devenu le terme qui désigne la vraie intelligence artificielle. Personnellement, je pense que, pour répondre un peu plus globalement, l'intelligence artificielle ce sont des mécanismes calculatoires qui vont traiter des données. Ces données peuvent être de toutes sortes, numériques, ou peuvent être symboliques au sens qu'on les a déjà prémachées, et donc on a donné des symboles. Mais il n'en demeure pas moins que ce sont des calculs. Des calculs qui, par définition, parce qu'on utilise une machine de Turing, vont être des calculs algorithmiques. Que les algorithmes soient explicites, c'est-à-dire écrits par des êtres humains, comme on fait classiquement des algorithmes, ou implicites, c'est-à-dire produits par des mécanismes d'apprentissage statistique – et je souligne statistique, c'est de ça qu'il s'agit –, qui produit un modèle qui finalement est capable de prendre de nouvelles entrées et de produire des sorties, mais ce faisant, ce modèle a intégré une sorte d'algorithme de traitement (mais qui n'est pas explicite, qui n'a pas été programmé explicitement), dans les deux cas, ce sont des calculs. Donc ces algorithmes et ces systèmes ne pourront jamais sortir du cadre de ce calcul, qu'il soit explicitement programmé ou implicitement programmé.

L'intelligence qu'on appelle « intelligence forte » exigerait qu'on puisse sortir de ce cadre. Un peu de manière comparable à ce que fait le cerveau humain. On n'a encore jamais prouvé que le cerveau humain est une machine de Turing. Donc ça, c'est un point d'interrogation, évidemment. Je n'en sais rien et personne ne le sait. On fait comme si c'était le cas, pour beaucoup. Je veux dire, quand on parle d'intelligence artificielle, on fait comme si le cerveau humain était une machine de Turing, voire une machine de Turing un peu complexe, parce qu'elle ne comprend pas un seul algorithme, une seule entrée, etc., mais c'est quand

même une machine de Turing sur le fond. Donc, pour arriver à l'intelligence artificielle forte, il faudrait que l'on ait dans la machine des capacités de traitement qui puissent sortir du cadre de ce qui a été explicitement appris, ou explicitement programmé. Ce n'est tout simplement pas le cas. Donc je ne vois pas du tout comment on pourrait aboutir à une intelligence artificielle forte avec la définition qui est communément admise, c'est-à-dire une intelligence qui soit capable de ne pas être focalisée sur tel ou tel type de données, ou tel ou tel domaine, et qui soit, disons, suffisamment générale pour être comparable à l'intelligence humaine. Tu es muté (anglicisme).

Mehdi Khamassi [26.29]

Voilà ! (rires) C'est intéressant parce qu'effectivement dans ce débat on a aussi la question des définitions qui contraint l'angle d'attaque, et en même temps mon impression est quand même que, dans la communauté de chercheurs actuellement en robotique autonome ou en intelligence artificielle, ça paraît presque comme une évidence que d'ici un certain nombre d'années, on ne sait pas combien, on pourrait atteindre en tout cas des niveaux d'intelligence sur la machine qui soient comparables à l'humain – donc ce qu'on peut appeler « human-like intelligence », ou des termes comme « artificial general intelligence » –, ou même une intelligence qui dépasse l'humain. Au-delà des termes même, il y a quand même un certain nombre de travaux, et puis des choses sur lesquelles on a collaboré aussi ensemble : comment organiser, au sein d'un agent, une architecture cognitive informatique, qui va combiner plusieurs fonctions cognitives, de la perception, de la mise en mémoire, de l'analyse de la connaissance qui peut y avoir en mémoire pour ensuite délibérer, et réagir à une situation présente, proposer une action, éventuellement apprendre sur la base des conséquences de l'action, réfléchir sur ce que l'on a fait. Avec cette réflexion sur les architectures cognitives, on peut se demander : qu'est-ce qui pourrait constituer une barrière, peut-être sur le long terme, à permettre (1) un haut degré d'autonomie de la machine, et puis (2) une adaptation à des situations qui ne sont pas prévues par l'homme, donc sortir du cadre du calcul qui a été prédéfini, tel que tu le décris ?

Différences avec le cerveau humain

Raja Chatila [27.22]

Évidemment, l'existence matérielle du cerveau humain ou animal montre qu'il est possible qu'il y ait une entité qui soit capable de raisonner, au sens humain du terme, de prendre des décisions, d'être autonome, d'avoir une capacité de réflexivité, etc. Donc il y a une preuve d'existence en quelques sortes. Mais le fonctionnement de ce cerveau n'est pas compris. On n'a pas encore élaboré un modèle qui puisse se traduire effectivement par une formulation qui pourrait être effectivement mise en œuvre sur une machine, donc qui pourrait être calculatoire. Ça ne veut pas dire qu'on ne pourra jamais le faire. Je ne dis jamais « jamais ». Mais je pense que tant qu'on n'a pas réalisé un système artificiel qui puisse mettre en œuvre des mécanismes que je ne connais pas encore, et qui ne sont pas forcément des mécanismes imitant le cerveau, mais des mécanismes qui puissent effectivement avoir cette réflexivité, avoir cette capacité d'auto-évaluation, avoir cette capacité de métaraisonner, i.e., de raisonner sur soi-même, on n'aura pas accompli le pas nécessaire pour franchir la frontière entre une machine au sens « machine de Turing », et une entité, j'évite d'utiliser le mot « machine », qui est le cerveau [humain]. Je ne sais pas comment on y arrivera. Mais ce qui est quand même clair, c'est que le chemin que l'on prend actuelle-

ment avec l'intelligence artificielle n'est pas ce chemin-là. On en reste au niveau syntaxique. On en reste au niveau du traitement de données de plus en plus importantes, de plus en plus massives, pour essayer de manière mécanistique, à partir de ces données, de produire des résultats ou des comportements.

Donc la question architecturale que tu poses est une question fondamentale, parce que c'est probablement dans l'organisation des systèmes que se trouve peut-être l'un des secrets, l'une des voies, pour essayer de comprendre comment on peut raisonner sur ses propres actions, les évaluer, comment on peut avoir cette sorte de métaraisonnement que je pense essentielle pour parler d'une intelligence plus proche de celle de l'être humain. Et de nouveau pour moi ça ne peut se produire que comme résultat, à la fois un résultat mais aussi une condition, de l'interaction de ce système avec son environnement. Parce que c'est l'unique manière de comprendre la sémantique de l'environnement ; c'est d'interagir avec lui. Donc pour le moment je ne sais pas comment faire. Comme tu le sais, on a travaillé sur des projets qui se voulaient aller dans ce sens. Mais on était très loin d'aboutir. Et l'idée même d'avoir essayé de mettre dans un système des mécanismes d'apprentissage différents, par exemple, et des choix pour passer d'un apprentissage habituel à un apprentissage basé sur des modèles ou des objectifs, inspirée en cela des modèles des neurosciences, montre qu'il y a peut-être une voie là. Mais cette voie est restée pour moi très limitée par le fait qu'elle était préétablie. Il manquait singulièrement de plasticité, une capacité d'évolution, une capacité de choix de ce qu'on appelle le métacontrôleur, de choix plus informés, plus approfondis. Je ne sais pas quel terme il faut utiliser, car il n'y a pas de terme pour désigner ça qui soit correct.

Donc oui, il y a des questions architecturales. Et bien sûr quand on regarde le cerveau on voit qu'il y a des réseaux de neurones dont le comportement, le fonctionnement est un peu semblable aux réseaux de neurones de l'apprentissage profond, on voit qu'il y a des architectures qui sont tout à fait différentes. Donc il est possible de s'inspirer de cela pour construire des systèmes qui dépassent les limites de l'IA faible. Mais je n'en ai pas la moindre idée. Je ne sais pas comment faire. Tout le reste n'est que projection dans un futur incertain. Et je ne sais pas si c'est la bonne voie. Je suis très hésitant dans ma réponse. La question est très importante. Mais on n'a pas de piste réelle pour essayer d'y aboutir. Donc ce sont des considérations un peu non fondées, non scientifiques. C'est difficile.

Mehdi Khamassi [26.29]

C'est vrai que dans les débats actuels, beaucoup de gens sont tentés de ramener la question à celle d'une IA avec une super intelligence qui arriverait à un moment donné dans la société. Or, ce que je comprends de ce que tu dis, c'est que ce n'est pas forcément là le cœur des questions qu'on peut se poser sur l'impact sociétal de l'IA, sur les questions d'éthique, sur ce que l'on en fait. Ce n'est pas sur ça qu'il faut réagir pour l'instant, qu'il faut pouvoir réguler et organiser, parce que c'est encore au stade de recherche et qu'il faut d'abord clarifier encore tout ça et mieux fonder tout ça.

Quelle échéance pour une super intelligence ?

Daniel Andler [34.41]

Si je peux, juste en complément. Moi je suis tout à fait enchanté de la réponse de Raja, qui me conforte dans l'idée que je pouvais m'en faire. Mais je ne suis pas un spécialiste qui fait le boulot. Je regarde de loin. Mais je voulais quand même lui demander : comme il doit certainement le savoir, il y a eu une enquête auprès des centres ou des grands chercheurs en IA, en leur demandant à quelle échéance

ils pensent qu'il y aura une super intelligence, ou du moins une « human-level intelligence ». Et les réponses s'étageaient entre « dans 20 ans », « dans 30 ans », « dans 50 ans ». Et ce qu'en ont conclu les journalistes, c'est qu'on ne sait pas si c'est 20, 30, 50 ou 100, mais enfin, ça va arriver fatalement. C'est quand même étrange d'une certaine façon que lorsqu'ils répondent à des questionnaires, ou qu'ils répondent à des journalistes, beaucoup de tes confrères se laissent quand même aller à faire des prédictions qui, au fond, ne sont pas vraiment sérieuses. C'est purement intuitif d'après ce que tu dis.

Raja Chatila [35.56]

Alors, il y a plusieurs catégories de chercheurs et de réponses. Il y a ceux qui confondent, de nouveau, vitesse de calcul, calcul, taille mémoire, avec intelligence. Ce sont les tenants de la singularité, disons plus ou moins, qui ont choisi une courbe exponentielle judicieusement, en choisissant les points par lesquels elle pourrait passer, s'inspirant de la loi de Moore qui n'a rien de scientifique, bien sûr, pour dire « voilà, en 2045, le 23 septembre [mettons], on atteindra un point où les ordinateurs, la machine, pourra dépasser le cerveau humain. » Évidemment, ils n'ont aucune idée de comment. Le calcul, la capacité de calcul, et d'ailleurs celle-ci existe déjà qui dépasse le cerveau humain dans plusieurs domaines, mais la capacité de calcul toute seule n'est pas suffisante évidemment. À la limite, je dis que puisque c'est comme ça, je prends mes 100 milliards de neurones et je les jette sur une table, comme ça, éparpillés. Ça ne fait pas mon cerveau. Donc le fait de dire ça n'a pas de sens. Mais les tenants de la singularité ont aussi un programme, disons, des intérêts pour dire ça.

Et puis il y a les chercheurs plus honnêtes, je dirais, qui pensent que puisque c'est notre objectif de faire des systèmes d'intelligence artificielle, on va y aboutir un jour. D'une certaine façon, c'est une honnêteté de dire « je travaille sur un programme qui un jour va aboutir à quelque chose ». Mais ils n'en ont pas la moindre idée. En réalité, je pense, d'ailleurs, que pour la plupart, ils n'ont pas un tel programme de recherche, mais sont plutôt focalisés sur certains éléments.

En gros, l'idée c'est qu'il y a un chemin, qui est sans doute escarpé, mais le sommet de la montagne existe : c'est le cerveau humain. On y arrivera bien un jour. Peut-être. De nouveau, je ne dis jamais « jamais ». Mais je pense qu'il est très probable qu'il faudra attaquer ça par une face nord et par un passage qui n'a pas encore été découvert. Et puis il y a beaucoup de gens qui répondent parce qu'on leur pose la question. Les journalistes sont assez doués pour sous-tirer des réponses même quand on ne veut pas répondre. Mais je n'ai absolument aucune idée sur comment on peut atteindre cela, ni quand, pour être très honnête.

2. Action conjointe entre un robot et un humain

Mehdi Khamassi [0.12]

Pour revenir sur les architectures cognitives, qu'on a évoquées précédemment, il y a une question de Jacopo Domenicucci, qui aurait beaucoup aimé participer mais qui n'a pas pu être là, qui s'intéresse beaucoup à tes travaux sur les architectures cognitives, et notamment dans quelle mesure cela permet l'action conjointe dans l'interaction entre un robot et un humain. Je lis sa première question : quelle différence y a-t-il dans la conception technique mais aussi dans les défis d'intégration sociale et éthique, entre les agents artificiels avec lesquels nous pourrions entreprendre réellement des actions conjointes, et des agents artificiels avec des niveaux d'intelligence (ou d'autonomie, d'apprentissage) élevés mais incapables de ce type de coordination avec l'humain ?

Raja Chatila [0.56]

L'action conjointe implique la capacité pour l'agent de comprendre l'être humain (j'utilise là le terme comprendre dans un sens faible). Si on ne peut pas comprendre, c'est-à-dire interpréter ce que fait l'autre, avoir un modèle de comportement de l'autre, avoir une théorie de l'esprit qui permette d'anticiper et de se mettre à la place de l'autre, on ne peut pas faire vraiment faire de l'action conjointe. Il faudrait que le système qui interagit avec les êtres humains puisse disposer d'un tel modèle, pour avoir cette théorie de l'esprit de l'être humain : à la fois un modèle physique de ce que l'être humain est capable de faire, pas seulement ce qu'il pense, donc un modèle physique géométrique, un modèle des systèmes de perception humains, un modèle qui lui permette d'agir de façon à ce que l'être humain le comprenne et puisse aussi coopérer avec lui. Si par exemple je veux tendre un objet à un être humain, je vais placer cet objet dans l'espace de travail de l'être humain que je connais, de manière à ce que l'être humain puisse le voir et le prendre. Je ne vais pas le poser au-dessus de sa tête. Tout ça veut dire des modèles, des modèles calculatoires qui peuvent être appris ou préprogrammés pour partie, des modèles de l'être humain qui sont nécessaires pour cette interaction.

Mais il faut aussi avoir un modèle des comportements, des actions (« qu'est-ce que l'être humain est capable de faire ? »). Mais si on parle d'intégration sociale et éthique, il faut aussi avoir une représentation des valeurs humaines, une représentation des préférences, et le cas échéant, avoir même une interprétation contextuelle. En effet, les valeurs sont en tension les unes avec les autres. Parfois, il y a des priorités et selon la situation l'être humain pourra avoir des priorités différentes. Donc, comment comprendre cela ? C'est là qu'il y a une limite quand même à cette capacité d'intégration sociale. Mais c'est là aussi où il y a une nécessité que la machine puisse avoir une certaine capacité de raisonnement, d'élaboration de modèles de comportements, qui puisse tenir compte de cela. S'il n'y a pas cette capacité, évidemment les systèmes artificiels, les agents artificiels en question, ne pourront jamais véritablement se coordonner avec les êtres humains. Ils pourront aller explorer Mars tous seuls le cas échéant. Mais ils ne pourront pas avoir cette action conjointe. L'action conjointe exige un minimum de compréhension de ce qu'est un être humain, de ses capacités, de ses préférences, de ses valeurs. C'est contextuel et c'est donc beaucoup plus compliqué qu'on pourrait le penser.

Nous-mêmes, nous avons travaillé sur ces sujets-là dans un cadre très limité. Il faut bien le dire, car on a souvent tendance à trop généraliser à partir de nos propres travaux. Nous y travaillons dans un cadre où nous avons simplement essayé de démontrer que certains éléments que je viens d'exposer sont possibles et semblent aussi être validés par des études psychologiques sur les êtres humains. Nous employons des architectures qui comportent plusieurs niveaux : un niveau de décision, un niveau d'exécution, de contrôle, et un niveau fonctionnel des capacités de base. Cette architecture cognitive artificielle devrait pouvoir rendre compte de ces différents éléments.

Lien avec les architectures cognitives humaines

Mehdi Khamassi [5.54]

Ça permet justement de rebondir avec la deuxième question de Jacopo, qui justement se demande quel rapport tu vois, au niveau des stratégies de recherche, entre l'étude des architectures cognitives humaines et artificielles (par exemple robotiques). Est-ce qu'il y a un éclairage réciproque ou bien pour l'instant il s'agit simplement d'un rapport d'imitation de l'humain par la robotique ?

Raja Chatila [6.08]

Ah non, bien sûr qu'il y a un rapport réciproque. En robotique on va élaborer des méthodes computationnelles qui pourraient très bien, pour certaines d'entre elles, être utilisées pour expliquer certains phénomènes ou certains comportements dans des architectures cognitives humaines. La formalisation et la faisabilité, c'est-à-dire la possibilité de mise en œuvre, font que le robot devient un outil pour montrer à la fois la possibilité théorique au moins, et éventuellement au-delà, les limites de certaines capacités humaines. Pourquoi ? Parce que dans le cerveau il y a quand même des capacités calculatoires : du raisonnement bayésien par exemple ; un filtrage de Kalman, qui est aussi une forme de traitement bayésien. Tout ça peut être mis en œuvre dans des machines, dans des robots, et permettre de valider ou d'inspirer la recherche sur les architectures cognitives humaines.

En même temps, dans l'autre sens, bien évidemment, n'oublions pas que notre programme n'a qu'une seule source d'inspiration, qu'un seul modèle : ce sont les architectures cognitives humaines. On ne les connaît pas bien, qu'on ne maîtrise pas bien, mais qui peuvent évidemment inspirer, et qui le font très souvent, des choix que nous faisons dans la conception de systèmes artificiels. Pour moi ce n'est pas de l'imitation. Je préférerais le terme inspiration. D'abord ce ne sont pas les mêmes mécanismes nécessairement. Et ce ne sont pas non plus les mêmes résultats qu'on va obtenir. Les concepts pourraient être similaires, mais il ne s'agit pas d'imitation. On voit très rapidement, d'ailleurs, que quand on essaie de mettre en œuvre certains concepts sur la machine, sur le robot, on va devoir faire des choix qui ne sont pas nécessairement les mêmes que ceux qui seraient faits si on voulait imiter strictement ce qui se passe dans les systèmes naturels.

Le cerveau humain et les robots suspendent-ils la tâche en cours de la même manière ?

Daniel Andler [9.12]

Est-ce que je peux intervenir ? Je n'y avais pas pensé mais tout à coup j'ai envie de poser une question sur justement quelque chose qui est modélisé chez l'homme. Je n'ai pas du tout la compétence pour savoir si le modèle est plausible, mais j'avais suivi les travaux d'Étienne Koechlin, qui travaillait sur la question de savoir comment est-ce le cerveau fait pour suspendre l'exécution d'une certaine tâche, pour veiller à exécuter une autre tâche qui par exemple est plus urgente, ou pour des raisons d'attention, et de revenir à la tâche d'origine. Ça, je suis sûr que c'est un problème qui se pose aussi en robotique. Un robot peut être en train d'essayer d'explorer quelque chose, mais tout à coup il y a un problème mécanique, un caillou ou je ne sais quoi, donc il va s'occuper de ça et puis revenir à la tâche initiale. Est-ce qu'il y a des mécanismes en robotique qui permettent au robot autonome de se débrouiller dans une situation comme celle-là ?

Raja Chatila [10.10]

Alors, tout à fait ! C'est une excellente question parce que ça me rappelle beaucoup de choses qu'on a faites dans le passé dans le domaine de la robotique. C'est le fameux problème entre les comportements et les actions délibératives, orientées vers un but, et les actions réactives qui sont nécessaires quand il y a un événement qui se passe dans l'environnement et dont il faut tenir compte. Il

existe de multiples exemples, ne serait-ce qu'en termes de déplacement : je veux aller quelque part, et puis il y a un événement qui va m'empêcher d'y aller, par exemple un obstacle qui se présente devant moi. On simplifie ça en disant « évitement d'obstacle ». Non ! Il ne s'agit pas d'évitement d'obstacle parce que je pourrais être obligé de choisir un chemin complètement différent. Et puis chemin faisant, comme le robot doit être conçu pour effectuer plusieurs objectifs, et pas seulement un seul, moi robot je peux m'apercevoir que je suis arrivé à proximité d'un lieu où je pourrai exécuter un autre objectif que j'avais dans ma pile d'objectifs. Est-ce que je le réalise, quitte à revenir ensuite à mon objectif initial ? Tout ça, ce sont des questions qu'on a rencontrées depuis très longtemps en robotique. Des mécanismes pour faire cela existent bien sûr. À la limite on pourrait même dire que c'est quelque chose de base en informatique. L'ordinateur multitâches, le système d'exploitation multitâches, il gère finalement en permanence la capacité de réaliser plusieurs tâches en partageant le temps, ou bien en mettant des priorités, ou en faisant des interruptions. Donc il y a plusieurs mécanismes de base pour cela. Est-ce qu'il faut interrompre une tâche pour en prendre une autre ? Est-ce qu'il faut l'insérer, donc essayer d'optimiser finalement l'ensemble des tâches ? Il y a beaucoup de mécanismes et beaucoup de méthodes pour gérer ces problématiques-là.

Très sincèrement, je ne suis pas du tout certain que s'apparente à ce qui se passe dans le cerveau ce qu'on fait pour gérer des systèmes multitâches dans les machines, dans les ordinateurs de manière générale, et dans les robots. Je ne le pense pas d'ailleurs. Mais pour partie oui. En robotique en tout cas, on a vu énormément de travaux pour réaliser des architectures qui soient à la fois délibératives et réactives en même temps.

D'ailleurs, je reviens à Rodney Brooks, que je mentionnais tout à l'heure, parce que dans son approche comportementale, on se retrouvait à donner la priorité à la réactivité par rapport à la délibération. Dans son esprit, plusieurs couches qui étaient toutes bouclées sur l'environnement, au bout du compte pouvaient aboutir, selon lui, aboutir à faire des comportements orientés vers des buts, simplement par des mécanismes d'inhibition ou de désinhibition ; une couche peut en inhiber une autre, mais l'autre va pouvoir réagir étant donné qu'elle est plus bouclée avec l'environnement réel, par exemple lorsqu'elle doit être actionnée pour tenir compte d'un événement imprévu. Donc ces mécanismes d'inhibition finalement étaient un peu complexes, et en réalité on n'a jamais abouti à faire une telle architecture qui soit capable réellement d'être « mise à l'échelle », c'est-à-dire d'effectuer des comportements complexes ; c'en est resté à des comportements limités.

Mais oui, ce problème de gestion des comportements délibératifs et de comportements réactifs est un problème central dans la conception d'architectures robotiques et il y a énormément de travaux dessus, qui ne s'apparentent pas tous à une inspiration des neurosciences.

Mehdi Khamassi [15.06]

D'ailleurs, il y a des petites différences aussi, même s'il y a des processus « d'embranchement », comme les appelle Étienne Koechlin, qui consistent à revenir à un but précédent ou autre, dans ses travaux notamment les plus récents, il montre des limites des capacités de l'humain à faire des embranchements entre plus que deux buts objectifs. Et puis on connaît aussi les limites de mémoire de travail en psychologie humaine, qui sont des contraintes qui ne se posent pas sur la machine. Donc pour le coup, ça aussi ça peut aboutir à des différences.

Daniel Andler [15.32]

J'ajoute aussi que la cognition vieillissante se caractérise, semble-t-il, par une difficulté à revenir à la tâche d'origine. Voilà, c'est un de nos handicaps parmi d'autres. (rires)

Raja Chatila [15.49]

Alors est-ce que c'est un problème de mémoire ou c'est un problème d'incapacité à revenir ? Est-ce que ça veut dire qu'on a un registre qui est de plus en plus limité pour pouvoir repartir ? Je fais une comparaison avec une machine, désolé. Ou est-ce que c'est quelque chose qui est perdu ?

Daniel Andler [16.18]

Je crois que c'est autre chose : ça vient du fait, moi j'ai vu ça trop vite et j'oublie beaucoup (rires), mais c'est les travaux de Patrick Lemaire à Marseille, notamment, qui montrent qu'en fait que les stratégies cognitives en général sont assez différentes. J'ai un peu oublié pourquoi. Mais ceci explique que dans beaucoup de tâches, ça marche très bien, et dans d'autres les questions d'interruption ne sont pas facilement résolues. Je ne sais plus très bien pourquoi. Mais tout cas il y a une grosse différente, semble-t-il, dans les stratégies de résolution de problèmes.

Vers une IA autonome ou plutôt un attelage humain-machine ?

Mehdi Khamassi [16.58]

Avant de passer à une autre partie de questions qui serait plus sur le déploiement de l'IA dans la société, des questions d'éthique et d'intelligence collective, est-ce que peut-être d'autres membres du groupe voulaient poser des questions sur ces aspects fondamentaux de recherche ?

Daniel Andler [17.15]

Alors je pose la question, une question assez évidente à laquelle Raja peut répondre très vite s'il le souhaite : j'ai essayé en faisant un petit peu le tour de recherches en IA ces derniers temps et je crois qu'il assez clair qu'il y a, disons, une direction qui veut vraiment pousser vers l'IA la plus autonome possible, confier complètement une tâche à l'IA, et une autre [direction], tu as parlé d'Englebarth à un moment de ton audition, Englebarth ou Jordan par exemple, qui parlent vraiment « d'intelligence augmentée », d'un attelage homme-machine, et que c'est ça qui a beaucoup d'avenir, et c'est là-dessus qu'il faut vraiment insister, et arrêter de s'obnubiler sur une intelligence artificielle qui serait autonome et qui ferait tout en l'absence de l'intervention humaine. Est-ce que tu crois qu'il y a un choix à faire ? Est-ce que tu as une préférence entre une orientation vers un attelage homme-machine d'une part, et une orientation vers une IA ou une robotique autonome ?

Raja Chatila [18.16]

La réponse courte, parce que tu m'as demandé une réponse courte.

Daniel Andler [18.22]

Tu peux la faire longue. (rires)

Raja Chatila [18.25]

Je vais élaborer après. La réponse courte c'est l'attelage humain-machine, évidemment. C'est-à-dire la préservation de la décision humaine, aidée par la machine. Alors, la petite entorse que je vois à ça mais qui n'en est pas vraiment : par construction, mais de manière inévitable, il faut laisser la machine agir complètement. Par exemple, si on parle d'une machine qui est bouclée sur l'environnement à une vitesse telle que l'intervention humaine n'est pas possible. À ce moment-là, la machine devient inutile. Par exemple, la voiture à conduite automatisée, typiquement, si elle est complètement développée et utilisée, on ne voit pas très bien l'intervention humaine puisque c'est contradictoire avec l'idée que la voiture conduit toute seule. À ce moment-là c'est à un autre niveau que se passe l'attelage humain-machine. L'homme ne peut pas intervenir directement dans le fonctionnement ni dans la décision de la conduite. C'est dans un cadre plus global que les choses devraient avoir été définies. Mais à part cet exemple ou d'autres de même catégorie, disons (et mon exemple n'est possible que pour une certaine raison, je vais y arriver tout de suite, qui explique pourquoi il vaut mieux toujours l'attelage), il est clair que, tout simplement, la machine ne comprend pas ce qu'elle fait. Cette absence de sémantique fait que les décisions de la machine peuvent être complètement inadaptées, erronées par rapport à la situation réelle, au contexte. Je parle de contexte sous forme d'interprétation sémantique. Il n'y a que l'homme qui peut donner ce sens. Et la machine peut l'aider à prendre ses décisions. Ceci est vrai dans beaucoup de domaines.

Je prends un autre exemple qui est celui de l'imagerie médicale, où on dit aujourd'hui que les systèmes de deep learning sont capables de faire des traitements d'image très performants et d'interpréter par exemple sur ces images s'il y a des tumeurs cancéreuses ou non cancéreuses, etc. Est-ce qu'on doit laisser la décision finale à la machine ? Évidemment non. Et ce n'est pas seulement parce que le domaine est risqué et qu'il y va de la vie des patients. C'est aussi parce que, par définition, la machine ne va voir que ce sur quoi elle a été entraînée. Elle a été entraînée sur certains types de tumeurs, donc elle va peut-être voir et distinguer des tumeurs, avec une certaine précision. Mais la machine peut ne pas voir quelque chose qui est là, dans l'image, sur lequel elle n'a pas été entraînée, et que seul le spécialiste verrait. Et la machine peut aussi se tromper avec une grande certitude, c'est-à-dire donner des résultats avec une très grande précision, mais qui sont complètement faux. Tout simplement parce que le mécanisme même d'apprentissage profond qui est utilisé est un mécanisme qui va chercher des régularités dans l'image, et ces régularités peuvent ressembler à quelque chose qui est une interprétation complètement erronée par rapport à la situation réelle qui est dans l'image qui est en train d'être traitée. Donc l'intervention humaine est essentielle parce que le sens et le contexte ne peuvent être apportés que par l'être humain.

Pour la voiture en question, c'est bien parce qu'on est dans un domaine très délimité, qu'on a réduit considérablement la sémantique, et qu'on peut accepter ce comportement où la machine prend la décision d'elle-même.

3. Sur le déploiement de l'IA dans la société

Mehdi Khamassi [0.12]

J'aimerais qu'on bascule maintenant sur des questions de déploiement de l'IA, et son implication dans la société à différents niveaux. On sait qu'il y a tout un mouvement de mode en ce moment,

de plus en plus d'entreprises veulent avoir leurs chercheurs en IA. Il y a par ailleurs toujours des recherches dans le milieu universitaire, mais il y a néanmoins un certain nombre de chercheurs universitaires qui partent dans des start-up, en entreprises. Les géants du Web se positionnent, intéressés à faire beaucoup d'applications. Et puis les États tentent tant bien que mal de réguler. Une question un peu globale, un peu méta, pour rentrer direct dans le vif du sujet : est-ce que tu penses que la façon dont les États, l'Europe, mais aussi les sociétés savantes essaient de gérer cette poussée de l'IA, relève d'un processus d'intelligence collective ? Et si non, qu'est-ce qui selon toi devrait être amélioré ? Est-ce que le travail que font les comités d'éthique suffit ? Est-ce qu'il manque quelque chose à tout ça ?

Les comités d'éthique suffisent-ils ?

Raja Chatila [1.12]

C'est une question difficile. Ça m'amène à soulever la question de l'expertise. Parce que les comités d'éthique sont en général composés par des experts. C'est-à-dire des gens qui ont une certaine connaissance d'un domaine, soit le domaine de l'intelligence artificielle en l'occurrence, pour ce qui nous concerne, soit par exemple ce sont des juristes, des sociologues, en tout cas plutôt des personnes qui sont des expertes dans un domaine donné. Donc il manque dans les comités d'éthique la perception de l'homme de la rue, le citoyen plus exactement. Et donc la question ici touche à comment impliquer les citoyens dans cette réflexion éthique ? Je n'ai pas de solution miracle. Mais je crois que c'est ce qui peut manquer aux comités d'éthique.

Les comités d'éthique parfois font des auditions, ils font des consultations, notamment avec des questionnaires, etc., pour essayer de sentir un peu comment l'ensemble des citoyens ressentent les choses. Mais ce n'est pas suffisant, car le citoyen ne participe pas ainsi à la réflexion elle-même. Je pense qu'on pourrait utilement améliorer l'appréhension de ces questions par la société si on avait un moyen d'avoir des comités d'éthiques, disons d'experts, qui travaillent et incluent des citoyens. Je ne sais pas comment choisir ces citoyens. Peut-être au hasard, après tout. En tout cas ces citoyens pourraient apporter un éclairage différent. Évidemment cela ne se passera pas de la même manière. On ne pourra pas utiliser toutes les mêmes démarches, mais l'expérience a montré que des groupes de citoyens sont tout à fait capables d'appréhender des sujets du moment qu'on leur donne les informations et connaissances nécessaires, et d'apporter un éclairage tout à fait neuf à ces questions-là. Voilà pour ma réponse à une partie de ta question sur les comités d'éthique.

En même temps, les comités d'éthique, qui vont avoir une démarche philosophique et un peu technique, ont autre chose qui leur manque. Ou plutôt qui ne leur manque pas mais qui est inhérente à leur définition : ce sont des experts normalement indépendants, ou en tout cas ils devraient l'être. Ils ont donc une posture d'indépendance pour aboutir à des avis, des opinions, des propositions, qu'ensuite les décideurs politiques vont éventuellement prendre en compte ou pas. Le problème là aussi, c'est qu'on a un hiatus entre les décideurs politiques et les experts. Parce que les décideurs politiques, comme on le sait, ne lisent souvent pas ce qu'on leur donne. Ils veulent quelque chose de résumé, d'actionnable comme on dit en anglais. Et ils n'ont pas non plus la connaissance ou la formation nécessaire pour comprendre les tenants et les aboutissants. Donc là il y a un autre côté de la question qui est : comment les comités d'éthique doivent-ils travailler avec le politique ? Est-ce que le format « Avis », et puis le politique en fait ce qu'il veut, est le bon format ? Je pense que c'est quelque chose

qui est limité aujourd'hui ; on en voit les limites. Et il faut trouver un mécanisme d'implication aussi parlementaire peut-être, ou des chargés de mission peut-être, en tout cas ça ne pourrait pas être des ministres. En tout cas, il y a un vide qui me semble important à combler.

Voilà donc les deux volets qui me semblent manquer aux comités d'éthique.

Régulations au niveau des États ou européen

Mehdi Khamassi [6.32]

Merci beaucoup, ça clarifie les choses qu'on pourrait améliorer de ce côté-là. Et je me demande du coup, à un niveau plus haut, au niveau des États ou au niveau européen, si les tentatives de régulation de cette poussée de l'IA te semblent aller dans la bonne direction. Est-ce que, pareil, il y a des choses qui manqueraient à ton avis ?

Raja Chatila [6.52]

Alors il faut bien comprendre que ce n'est pas l'éthique qui motive les gouvernements. Ce qui les motive, c'est l'économie. Donc quand la Commission européenne fait une position de réglementation, celle-ci est dans un cadre qui est celui de la politique économique de l'Europe. Cela s'applique pour tous les domaines évidemment. L'intelligence artificielle est ici considérée comme un domaine scientifique et aussi comme une technologie au service du développement économique. Or à partir du moment où on réfléchit sur le développement économique, il y a beaucoup de facteurs dont il faut tenir compte : d'une part, la liberté d'entreprendre des entreprises, donc les moyens pour que ces entreprises puissent être performantes, par la concurrence internationale. Une certaine forme de soutien et/ou de protection, bien que le mot « protection » ne soit pas à la mode actuellement dans notre économie. Et c'est ça qui vient finalement ensuite essayer de tenir compte de certains principes d'éthique et où la réglementation va s'exprimer. Il faut bien tenir compte de cela : la réglementation européenne n'est pas issue des travaux sur l'éthique du groupe d'expert ; elle n'est pas issue d'une volonté d'avoir des systèmes d'intelligence artificielle qui respectent les valeurs éthiques. Elle est issue d'une volonté d'avoir une économie européenne forte utilisant une technologie puissante.

Mais ce faisant, on va regarder le paysage international, la concurrence et en particulier les États-Unis et la Chine, car ce sont les deux grands acteurs dans ce domaine. Et on va voir quel est le positionnement européen là-dedans, compte tenu des forces et des faiblesses des entreprises européennes. Donc pour moi les choix de la règlementation européenne s'appuient d'une part sur des cadres réglementaires précédents, préétablis, sur la protection des données personnelles, sur le respect de la charte européenne des droits humains, qui sont propres à l'Union européenne, mais qui sont aussi partagés par d'autres pays. Aux États-Unis, dans la constitution, normalement on respecte aussi les droits humains. Mais ils n'ont pas la même approche en termes de libéralisme économique. La protection du citoyen européen est plus importante en termes législatifs qu'aux États-Unis, et elle est tout à fait différente en Chine. Donc il y a des systèmes politiques qui sont tout à fait différents. C'est pour ça je crois qu'il faut lire la proposition de législation européenne, qui est une sorte de législation minimale, c'est-à-dire qu'on interdit les choses qui sont vraiment choquantes pour les droits humains en Europe. Par exemple, la surveillance de masse et la notation des citoyens. Ça, on dit « non, ce n'est pas acceptable », et donc c'est interdit. La surveillance néanmoins reste permise dans certaines

conditions. On ne veut pas la déployer en temps réel. La reconnaissance faciale en particulier n'entre pas dans le cadre législatif qui doit être bien défini. La tension sécurité/liberté est gérée de cette manière-là, par la règlementation.

Ensuite, on ne veut pas légiférer d'une manière qui mette trop de contraintes sur les entreprises. Pour tout ce qui concerne l'utilisation de l'intelligence artificielle dans des domaines dits « à haut risque », car ils concernent la vie humaine, comme dans le domaine de la santé, des transports, on va prendre la réglementation de ces domaines-là, qui est déjà très exigeante (comme par exemple la réglementation sur les dispositifs médicaux, ou sur les logiciels utilisés dans les transports, etc.), et on va dire : « si on met de l'intelligence artificielle là-dedans, celle-ci doit continuer à respecter les dispositions législatives qui sont celles de ces domaines-là. Donc c'est une orientation vers les applications. Ce n'est pas une législation qui concerne l'IA en général.

Et puis la législation dit que pour les domaines où les logiciels ne concernent pas les domaines sensibles, en termes d'intégrité, de santé humaine, etc., donc sont à part, on va faire une distinction en deux catégories : une catégorie qui ne nécessite aucune réglementation, aucune contrainte, parce que c'est dans des applications dans des secteurs qui ne prêtent pas à conséquence ; et pour l'autre catégorie, on va demander une certaine transparence et une certaine certification par des tiers, au sens de la marque CE. Par exemple si on utilise des logiciels pour faire du recrutement, analyse des CV, etc., l'entreprise qui va les développer et les utiliser va devoir les faire certifier par un tiers pour montrer leur conformité avec un certain nombre de critères, en particulier l'équité, la non-discrimination, etc.

Problème de l'autocertification

Raja Chatila [14.30]

Je trouve que c'est relativement minimal, parce que l'autocertification, et la certification par des tiers – qui va certifier les certifieurs ? –, tout ça n'est pas très défini. Les domaines vont être définis d'une manière agile. C'est une bonne chose, car on ne peut pas savoir a priori quelles sont les applications qui vont poser problème. Mais on en a prédéfini quelques-unes. L'objectif est de respecter un certain nombre de valeurs qui sont dans la législation européenne, mais en embêtant le moins possible les entreprises. Voilà ma lecture très résumée de la législation européenne. Donc il ne faut pas non plus l'interpréter comme quelque chose qui soit très exigeant pour ces entreprises-là.

L'une des difficultés de cette approche qui est basée sur le risque, c'est de définir le risque véritablement, et non pas en disant « la santé est un domaine risqué » et voilà. Le risque est un concept qui est difficile à manier, avec des probabilités le cas échéant, des dangers possibles, qui peuvent se manifester ou pas. C'est quelque chose qui n'est pas bien défini et qui va sans doute faire couler beaucoup d'encre. Aussi, les mécanismes de vérification et de validation des logiciels en vue de leur certification sont quelque chose qui va sans doute être amené à être précisé. Et on va sans doute voir beaucoup de start-up qui vont commencer à travailler sur ce domaine-là, je pense. En même temps, globalement, je pense que c'est une bonne chose, quelque chose de positif par rapport au paysage international ; c'est-à-dire le laisser-faire total par exemple États-Unis. Ceci est à mettre en comparaison avec la règlementation sur les données personnelles, le RGPD, qui protège quand même le citoyen européen, et qui peut devenir un exemple pour d'autres pays, pour exiger que les systèmes d'IA qui sont déployés sur leur territoire respectent un certain nombre d'exigences de ce type-là.

Disparité des moyens publics/privés et guerre de l'IA

Mehdi Khamassi [17.01]

J'aurais une dernière question, un peu dans ces dimensions mais pas exactement les mêmes : finalement, est-ce que la disparité des moyens, des géants du web et des entreprises et programmes chinois d'un côté, et des laboratoires publics français et sans doute européens de l'autre, est-ce que c'est quelque chose d'inquiétant ? Y'a-t-il un espoir de rééquilibrage ? Et sinon, comment vivre la situation le moins mal possible ?

Raja Chatila [17.25]

Ceci est la question de la guerre de l'IA. Il s'agit d'une illusion, à mon avis, qu'il y a une course à l'IA, une course à l'armement. Cette vision qui consiste à dire qu'il y a une guerre de l'IA, et que la Chine et les États-Unis sont déjà bien installés, avec leurs grandes entreprises, les GAFA d'un côté, les Weibo et autres Tencent de l'autre, etc., est une vision qui n'est pas très bien fondée. Parce que d'abord, revenons à la réalité de l'informatique. Les grands systèmes d'exploitation sont américains. Qu'ils soient en logiciel libre, par exemple UNIX et ses descendants – LINUX est européen, mais c'est un descendant de UNIX ou s'y apparente –, ou qu'ils soient propriétaires, comme Mac OS, comme Windows, etc., ils sont américains. Il n'y a pas de système d'exploitation européen véritablement, à part des logiciels libres. Il n'y en a pas de chinois non plus. Quand on regarde les mobiles, les téléphones portables, c'est la même situation. Avec le partage du marché globalement entre IOS et ANDROID. À cause de la politique américaine, les Chinois ont été amenés à développer leur propre système d'exploitation pour les mobiles, parce que Huawei n'avait pas l'autorisation d'utiliser ANDROID. Je dirais que les dés sont pipés dès le départ, parce que les États-Unis maîtrisent les systèmes d'exploitation sur lesquels tournent tous les ordinateurs.

Maintenant, oublions les systèmes d'exploitation et regardons les ordinateurs eux-mêmes. Sur quoi fonctionnent ces ordinateurs ? Les chips sont produits par qui ? Le géant, c'est Intel, même s'ils sont fabriqués ailleurs, y compris en Chine d'ailleurs. Je pense que la place des États-Unis, sans même parler d'IA, ne serait-ce qu'en parlant de moyens informatiques, est bien évidemment prépondérante. Il n'y a plus d'informatique véritablement européenne depuis longtemps. On a essayé. Et puis pour différentes raisons on a abandonné. Mais c'est comme ça.

Les entreprises américaines sont installées en Europe. Il y a IBM France. Il y a des centres de recherche, de Facebook, Google, etc. Donc il y a une très forte interaction entre l'économie américaine de l'intelligence artificielle et l'économie européenne de l'intelligence artificielle. Cette situation fait qu'il est très difficile de dire qu'on va prendre notre indépendance totale et faire notre Airbus de l'IA, comme on l'a fait face à Boeing. Beaucoup de gens disent que c'est plus vraiment possible. Ce n'est pas parce qu'ils ont tellement d'avance. C'est parce qu'on très imbriqués déjà. N'oublions pas par exemple que le Health Data Hub a confié à Microsoft ses données, qu'IBM est vendeur d'énormément de solutions en IA pour beaucoup d'entreprises françaises, etc.

Donc c'est une illusion de croire qu'il y a l'Europe, les États-Unis et les Chine, et que l'Europe doit trouver sa place entre les deux géants. On est déjà du côté américain, pour dire les choses. Et les Américains ont tout fait pour empêcher les Chinois de pénétrer notre marché, avec la bataille contre Huawei en particulier, et ce n'était même pas dans le domaine de l'IA mais dans le domaine des Télécoms.

Dire qu'on doit choisir de s'allier aux États-Unis contre la Chine, ou quelque chose comme ça, c'est une fausse question. On est déjà dans le domaine. Et je pense que poser les choses en termes

d'Occident contre l'Asie est forcément faux, parce que ce qui intéresse les Chinois, c'est de vendre au marché européen et au marché américain. Il ne s'agit pas de développer des têtes nucléaires qu'on va stocker pour les utiliser un jour. Donc cette image de la guerre et de la course aux armements est complètement fausse. Elle est là pour nous amener à forcer une alliance politique, et non pas dans le domaine de l'IA, dans la bataille commerciale des États-Unis même contre la Chine. Je pense que cette question de la concurrence entre les États-Unis et la Chine pour trouver la place de l'Europe est un peu faussée par tout ce contexte.

Problème de souveraineté numérique

Raja Chatila [23.35]

Maintenant, il y a quand même un problème de souveraineté numérique. Pourquoi ? À cause de tout ce que j'ai dit tout à l'heure. Par exemple, quand la France a voulu développer une application sur le traçage pour le Covid, la France a fait un certain choix sur la manière de gérer les données. Et ce choix n'était pas accepté par Apple. Donc ils ont dit « non, vous ne pouvez pas utiliser le système d'exploitation d'Apple pour que l'application soit tout le temps en train de traiter des données, que le Bluetooth soit ouvert tout le temps ». C'est-à-dire qu'un des géants du numérique, et je ne parle pas d'IA, impose à un pays souverain, et pas le moindre, des solutions techniques, et refuse tout simplement de faire autrement. D'ailleurs, vis-à-vis du gouvernement américain, Apple a une attitude similaire, pour ce qui concerne la protection des données de ses utilisateurs.

Ces entreprises ont une place qui met en cause la souveraineté des États, pas seulement dans ce type de choix que je viens de mentionner, mais aussi dans le domaine des réseaux sociaux, à Facebook, Twitter, etc., qui sont complètement supranationaux dans les moyens de communication entre les individus, et qui bien sûr sont le canal de propagation de toutes sortes de propos racistes, de problématiques liées au harcèlement, etc., pour laquelle les États ont beaucoup de mal à les réglementer, même si parfois il y a des tentatives.

Pour moi, le problème est là. L'IA n'est qu'un outil dans ce domaine-là. N'oubliez pas que l'idéal, ou plutôt le comble d'une start-up est de se faire racheter par une grosse boîte. Donc quand une start-up se développe dans le domaine de l'IA et se fait racheter par Apple, ils sont très contents. Cette idée de concurrence dans le marché dans lequel nous sommes n'est pas réelle. Le problème réside plutôt dans quelle règlementation il faut mettre en œuvre pour qu'il y ait une certaine souveraineté démocratique face à des entreprises géantes. Pour moi c'est ça la vraie question.

Développer par exemple une entreprise comme OVH, qui s'engage à garder les données personnelles en respectant la règlementation européenne, etc., ça fait partie du sujet, parce qu'on veut protéger les données. Et les outils d'intelligence artificielle qui sont derrière, ça fait partie du sujet. Mais il vaut mieux poser les choses de cette manière-là, je pense, plutôt que dans l'idée d'une course mondiale où l'acteur européen a une troisième place misérable et ne se préoccupe du coup que d'éthique et de philosophie. Ce n'est pas ça le sujet.

Comme je le dis sur la règlementation, c'est une règlementation minimale pour protéger d'une certaine façon le citoyen européen, conformément à des règlementations existantes, mais qui ne donne pas à l'entreprise européenne un avantage positif particulier.

Je pense qu'il n'y a pas d'approche spécifique à l'IA. C'est une question beaucoup plus globale sur l'économie du numérique d'une manière générale dans le monde, je crois.

Formation et éducation en IA

Mehdi Khamassi [28.09]

Eh bien merci, je comprends tout à fait. Est-ce que dans tout ce qu'on a discuté aujourd'hui sur les problématiques liées à l'IA, il y aurait des questions qui te sembleraient importantes à souligner et qu'on n'a pas abordées ?

Raja Chatila [28.29]

La formation et l'éducation sont des sujets qui sont importants. Quand je parle de formation, je parle de formation de ceux qui vont développer de l'IA : des chercheurs, des ingénieurs, etc., qui devraient avoir une formation qui inclut les questionnements éthiques qui sont aujourd'hui posés, et pas seulement être focalisés sur la course au traitement de la donnée et à la réalisation des systèmes, et éventuellement à la création de la start-up qui va commercialiser le système. Là il y a un problème de formation. Il y a aussi un problème de formation pour les entreprises qui développent. Je ne parle pas seulement de formation dans les écoles d'ingénieurs et les universités, mais aussi de formation des décideurs de l'entreprise, pour qu'ils soient formés à la nature de la technologie et de ses implications, de formation des politiques, c'est-à-dire dans les écoles – l'ENA n'existe plus, mais je ne me souviens plus comment elle s'appelle maintenant –, en tout cas la formation dans les écoles d'administration envers ceux qui vont gouverner, qui vont être les décideurs, et qui parlent de technologies qu'ils ne connaissent pas. Donc là aussi la formation doit être à la fois technique et sur les questions d'éthique qui se posent. Et puis l'éducation, l'information et la formation du public pour prévenir toutes les illusions que les gens ont sur les capacités de l'intelligence artificielle, les notions très approximatives, la peur de la prise de pouvoir de l'IA, tout ça. Il faut sortir de ces illusions et donc la médiation scientifique, l'explication au public doivent être quelque chose qui mérite une réflexion. Comment faire ça ? Comment mettre ça en œuvre ? Comme on dit en anglais : AI literacy, par opposition au fait d'être illettré.

Ceci est un sujet qui n'est pas au centre des réflexions que vous menez, mais qui est quand même quelque chose qui du point de vue sociétal devrait être pris en compte : faire en sorte que globalement l'ensemble de la société comprenne les enjeux et les réalités de ce domaine technologique.

Réflexion philosophique sur l'être humain et la machine

Raja Chatila [32.08]

Ensuite, à l'opposé de l'éducation, il faut mener la réflexion philosophique, multidisciplinaire – celle que vous menez ici – sur l'Homme (l'être humain) et la machine. Parce que l'un des problèmes que je perçois avec le développement de l'IA, est la réduction de l'être humain à une fonction machiniste ; la perception de l'être humain comme étant une sorte de machine en partant justement de modèles mal compris, qui sont des modèles computationnels, des modèles de comment fonctionne l'IA, comment fonctionnent les systèmes, les ordinateurs. À partir de ces modèles, on projette d'une part des capacités humaines sur les machines, et on réduit les capacités humaines en même temps en automatisant le comportement humain. En fait, les machines fonctionneraient mieux si les êtres humains fonctionnaient

comme des machines. Donc on essaie de les rendre le plus automatisés possible. Et en même temps on projette des capacités humaines, émotions, conscience, etc., sur des machines, alors qu'elles ne possèdent évidemment pas ces capacités-là. C'est très facile de projeter, c'est inévitable. On anthropomorphise les choses. Mais dans le cas présent, quand les choses en question s'appellent « systèmes intelligents » ou « intelligence artificielle », ça prend une dimension différente. Donc cette réflexion sur la différence entre l'être humain et la machine doit être plus ciblée dans le débat.

Mehdi Khamassi [34.36]

Merci beaucoup, Raja, pour ton temps et pour ton éclairage. On va clôturer pour aujourd'hui. Et si on a d'autres questions, on échangera avec toi par la suite.

Raja Chatila [32.08]

Avec plaisir ! Merci en tout cas.

4. IA générative et grands modèles de langage

Mehdi Khamassi [0.12]

Bonjour Raja. Merci de nous accorder encore du temps. Nous sommes le 14 juin 2023, et on voulait faire un complément de l'audition qu'on avait faite avec toi l'année dernière pour TESaCo. Ceci en particulier parce qu'il y a un débat en ce moment très important, qui est suscité avec le déploiement des grands modèles de langage (les large language models en anglais), et de toute la question de l'IA générative, donc qui génère des contenus, du texte bien sûr, mais aussi de l'image, de la vidéo, qui deviennent de plus en plus difficiles à discriminer de ce que peuvent générer les humains. Certaines personnes disent même que nous sommes face à « des risques existentiels pour l'humanité ». Nous voulions donc de poser la question : est-ce que tu peux reposer les termes du débat selon ton point de vue et quels sont les enjeux ?

Raja Chatila [1.01]

D'accord. Merci d'abord de m'avoir proposé cet entretien. Je réalise en particulier le temps qui est très compté pour tout le monde, et pour moi en particulier. Pour aller tout de suite à ces questions-là, il faut juste poser la question de la définition.

Définition de l'IA générative et des Transformers

Raja Chatila [01:31]

Je ne vais pas définir l'IA en général, etc., on en déjà parlé. Mais que sont ces nouveaux systèmes et pourquoi ils provoquer cet engouement, ce choc qui dès la parution de ChatGPT, le premier jour

il y a eu un million d'utilisateurs, et maintenant on a à peu près deux milliards de connexions mensuelles. Ça repose d'abord, techniquement parlant, sur quelque chose qu'on appelle les GPT (Generative Pretrained Transformers). Ce sont des architectures de réseaux de neurones qu'on appelle les Transformers, qui ont été proposées dans un papier par Vaswani et collègues, qui étaient à Google Deepmind (parce qu'entre parenthèses, c'est maintenant dans ces grandes entreprises du numérique que les nouveautés, qui étaient autrefois du milieu académique, nous arrivent. Cela ne veut pas dire qu'il ne se passe rien au niveau académique, mais les moyens sont beaucoup moindres).

L'idée du Transformer était de remplacer en quelque sorte quelque chose qui était effectué par l'utilisation de réseaux récurrents, pour essayer d'analyser un contexte quand on interprète des textes. La différence entre les textes et les images, essentiellement, c'est que dans une image tous les pixels sont là et on peut faire des traitements, des convolutions, on peut passer d'une couche à l'autre du réseau en gardant les éléments qui ont été détectés dans l'image, et les réassembler. Dans le texte, pour interpréter un mot, il faut le plonger dans un contexte, dans la phrase qui précède, voire dans un texte beaucoup plus long qui précède. Donc il y a un problème de mémoire, pour mémoriser ce contexte, et un problème d'architecture pour les réinjecter pour l'interprétation du mot courant. Là où on utilisait pour l'interprétation de la langue naturelle les réseaux récurrents, avec une démarche qu'on appelle Long Short-Term Memory (LSTM), qui n'était pas suffisante. Donc l'idée de base du Transformer était d'utiliser une architecture qui permet la réinjection d'un contexte beaucoup plus long. Et cette idée était basée sur quelque chose de pas totalement nouveau : il s'agissait de décomposer les mots en éléments, ou sous-mots, qu'on appelle tokens, qui sont porteurs d'une certaine information, et dont la combinaison va produire des mots. Ces tokens sont transformés en vecteur de façon à ce qu'on puisse les traiter et calculer des distances entre les vecteurs dans un espace latent. Ceci permet de prédire le prochain mot dans une phrase ou dans un texte, prédire aussi un mot manquant. En d'autres termes, si j'enlève un mot, par corrélation on va essayer de trouver le mot qui aurait pu être là.

Tout ça a provoqué une avancée importante dans le domaine du traitement du langage naturel. Et on a vu depuis 2017 (l'apparition de ce papier de Vaswani et de ses collègues) beaucoup de progrès, beaucoup de résultats, beaucoup de publications qui s'appuient sur cette idée, sur cette architecture. Ceci a déclenché ensuite une sorte de course : l'idée est que si on préentraîne une telle architecture sur de très grandes quantités de données textuelles (je me focalise sur le texte pour le moment), dans un apprentissage non supervisé, alors on construit un énorme modèle, qui est tout prêt en quelque sorte, car il contient énormément d'informations, qui est tout prêt pour être utilisé ensuite, soit directement, soit en le réentraînant sur un corpus spécifique, et pour produire des textes générés (puisque c'est génératif), en exploitant toute cette masse de données sur laquelle il a appris par capacité de corrélations. N'oublions pas que ce sont des corrélations. Plus on a de tokens, donc de puissance représentative, et plus on peut produire des mots, donc du texte qui est long à partir d'un élément initial de texte. Dans GPT-4 par exemple, on peut produire 50 pages de texte, ce qui est énorme ! On peut ainsi stocker un contexte qui est énorme.

Agents conversationnels avec grands modèles de langage

Raja Chatila [07:25]

Cette puissance était connue et évoluait depuis 2017. OpenAI, qui est une entreprise relativement petite, dite sans profits, a produit et rendu public en 2020 un modèle qui s'appelle GPT-2, et tout de

suite ils l'ont retiré en disant que l'accès ouvert est trop dangereux. C'est intéressant que je le mentionne, car aujourd'hui on voit un peu le contraire. Ensuite ils ont introduit ChatGPT. C'est un agent conversationnel utilisant d'abord GPT-3, ensuite GPT-3.5 et GPT-4. Un agent conversationnel est un système qui a une interface d'utilisation avec l'utilisateur. Comme son nom l'indique, conversationnel veut dire qu'on utilise la langue pour interagir, écrite sur un clavier ou orale, en l'occurrence ici c'est écrit. Et cet agent va produire, répondre, interagir. Tous ces agents conversationnels sont connus depuis longtemps, posent un certain nombre de questions éthiques aussi. Mais bien sûr ces agents sont limités par le système de génération de langage qu'on met derrière. Mais dès lors qui est derrière. Mais dès lors qu'on met derrière un système de génération de langage aussi puissant que GPT-3.5 ou GPT-4, on obtient quelque chose qui est capable de mener une conversation beaucoup plus riche, si je puis dire. Or ce qu'a fait OpenAI, c'est de mettre ça à la disposition du public ; c'est-à-dire de le rendre ouvert, et tout le monde pouvait y accéder, c'est-à-dire le grand public, n'importe qui, les médias, etc. Évidemment, ça a provoqué un événement assez particulier. Je m'explique : tout le monde avait entendu parler de l'intelligence artificielle, mais peu de gens savaient qu'ils interagissaient avec un système d'intelligence artificielle lorsqu'ils faisaient une requête sur un navigateur ou cherchaient leur chemin sur leur téléphone. Car derrière ces actions, il y a des algorithmes qui relèvent de l'intelligence artificielle. Tout à coup, tout le monde peut interagir avec un système qui porte un nom, qui est annoncé comme un système d'intelligence artificielle extrêmement puissant et qui manie le langage. Or le langage, la langue, est connu depuis l'Antiquité comme étant le propre de l'Homme, de l'être humain. Cela provoque un choc, je pense, dans l'opinion publique et dans les médias, car ces derniers se sont emparés de cela d'une manière absolument inouïe, en mettant en avant les capacités de l'intelligence artificielle à manier la langue. C'est vraiment considéré comme le nec plus ultra, pense-t-on, de ce qui fait notre intelligence, notre capacité à articuler des idées, des discours. En plus, les discours produits ou les textes énoncés par ChatGPT étaient très bien rédigés, sans faute d'orthographe, grammaticalement corrects, donc tout à fait lisibles. Sur le fond, disons que ce n'était pas dans un style très avancé, néanmoins bien construit et même plus que ça, avec des informations. Autrement dit, cela apporte quelque chose à quelqu'un qui ignore un fait, une situation ou des concepts. Cela explique à n'importe qui des concepts qui peuvent être assez complexes, qui ne sont pas à la portée de tous. Et je pense que cela a changé de manière qualitative l'appréciation de tout le monde, y compris des chercheurs, sur l'avancement de ces systèmes et leurs capacités. Cette illusion d'intelligence a encore progressé et a atteint un palier important. C'est pourquoi nous en parlons, c'est pourquoi hier j'étais à France Culture pour une émission. Tous les jours, je reçois des appels de journalistes qui me demandent ceci ou cela. Les journalistes sont constamment intéressés par ce sujet.

Des risques existentiels pour l'humanité ?

Raja Chatila [12:27]

Alors, maintenant, extinction ou pas extinction ? La pétition précédente qui, en résumé, demandait une pause, ce que j'avais signé, et l'appel plus récent évoquant l'extinction, que je n'ai pas signé, pose effectivement question. Pour traiter cela de manière assez succincte, en mars dernier, une pétition a été lancée par The Future of Life Institute, peu importe, et a été signée par beaucoup de monde. Elle argumentait sur un certain nombre de problèmes posés par l'intelligence artificielle générative, mais sans parler d'extinction, mais en évoquant néanmoins une menace pour un certain nombre de valeurs de l'humanité. La pétition demandait une pause de 6 mois le temps de prendre des mesures de

gouvernance concernant ces systèmes. Personnellement, dans cette pétition il y a un certain nombre d'affirmations avec lesquelles je ne suis pas d'accord. Je pense qu'une pause de 6 mois ne suffira pas, car ça fait des années que nous travaillons sur les mécanismes de gouvernance de l'IA et cela ne se résoudra pas en six mois. De plus, pour moi, 6 mois c'est relativement négligeable comme temps. Mais je l'ai signée quand même. Pourquoi ? Parce que, dans l'engouement et l'enthousiasme ambiant, je pense qu'il était nécessaire de lever un carton rouge. Il y a quelque chose qui ne doit pas être considéré uniquement comme une simple progression, mais plutôt comme un « game changer » en anglais, quelque chose qui change les règles du jeu, ou le match. Il faut y prêter attention, d'autant plus que tout le monde s'était rué sur l'utilisation de ChatGPT, que ce soit le grand public, les médias ou même les entreprises. Celles-ci envisageaient de l'utiliser, par exemple, pour le service client, l'analyse de divers sujets, voire la production de programmes informatiques, puisqu'il est également capable de les générer. J'ai donc signé cette pétition, tout comme Elon Musk l'a fait, (tout le monde mentionne Elon Musk et personne ne mentionne Raja Chatila, peu importe, ce n'est pas le sujet). On n'est pas forcément d'accord avec tous les signataires d'une pétition. Et cela a provoqué le choc que j'attendais, donc je suis finalement satisfait d'avoir signé. Le choc a été suivi de réactions médiatiques que l'on peut observer aujourd'hui et d'une discussion un peu plus approfondie sur ce que sont ces systèmes.

Ensuite, il y a une deuxième affirmation provenant de AI Safety, une autre institution, qui évoque brièvement le risque d'extinction. Elle a été signée par des personnes que je respecte, comme Yoshua Bengio, Geoffrey Hinton, des gens tout à fait au fait de ce que sont réellement ces systèmes, et de ce qu'ils sont capables de faire ou de ne pas faire. Cependant, elle a également été signée par les concepteurs mêmes de ces systèmes, Sam Altman par exemple, le PDG d'OpenAI, et d'autres grands acteurs industriels de l'intelligence artificielle de la Silicon Valley. Au début, ils ne l'avaient pas signée, mais étant en désaccord avec le terme « extinction », et voyant qu'ils faisaient partie des signataires, je me suis posé des questions et je ne l'ai pas signée pour cette raison.

Tout d'abord, le terme « extinction » est utilisé au même titre que pour parler de la pandémie ou du risque nucléaire. Je ne pense pas que les systèmes d'intelligence artificielle posent un risque d'extinction. Ils posent de nombreux problèmes, que nous pouvons aborder concernant la démocratie, le lien social, la question de la vérité, etc. Mais l'extinction est un concept extrêmement fort. Plus fort que l'extinction de l'espèce humaine, je ne vois pas. Il y a bien sûr la destruction de la planète entière et de toute vie sur celle-ci, mais si nous sommes éteints, nous ne le verrons pas. C'est donc exagéré et cela rejoint toute la mythologie autour de la super intelligence et de ses conséquences. Ce n'est pas fondé, ce n'est pas expliqué évidemment. Le terme « extinction » ne convient pas. Et le fait que ceux qui produisent ces systèmes signent une telle pétition devrait nous interpeller. Pourquoi celui qui produit un système et le met sur le marché dit-il en même temps qu'il peut provoquer l'extinction de l'espèce humaine ? Par exemple, ceux qui produisent de l'énergie nucléaire et construisent des centrales nucléaires, ou ceux qui fabriquent des bombes atomiques, ne disent pas « attention, cela peut provoquer l'extinction de l'espèce humaine ». Ils affirment qu'ils vont contrôler l'usage, etc., et les risques vont être gouvernés et maîtrisés. Il y a donc quelque chose d'original là, et une contradiction flagrante entre produire un poison, si je peux utiliser cette métaphore, et affirmer en même temps que c'est un poison extrêmement dangereux. Pourquoi ne pas arrêter simplement de le produire dans ce cas ?

Et donc, je pense que ce qui est sous-jacent à cette signature – parce qu'en parallèle il y a d'autres déclarations sur la réglementation de l'intelligence artificielle, qui commence à prendre une certaine dimension, ce n'est pas qu'en Europe qu'on en parle, on en parle au G7, on en parle dans la Silicon Valley également –, et eux, les producteurs de ces systèmes commencent à parler de réglementation et de besoin de nécessité de réglementation. Donc, je pense que cette opération, c'est aussi pour se positionner dans la conception de la gouvernance ou de la réglementation internationale de ces

systèmes, afin d'influencer de quelle réglementation nous parlons, et finalement d'être des acteurs de cette réglementation, au même titre que les gouvernements et les États. Donc, finalement, évidemment, il s'agit d'essayer de contrôler les différentes orientations. Je vais m'arrêter là parce que la question était importante, mais il était aussi important de développer un peu la façon dont ça s'est passé, pourquoi il y a eu ces appels, et finalement où on en est aujourd'hui. Aujourd'hui est un jour particulier, parce que le Parlement européen vote aujourd'hui sur son texte du règlement européen.

Hubris des entrepreneurs annonceurs de l'extinction

Daniel Andler [20.06]

Je voudrais poser à Raja deux questions. Une sur ce qu'il vient de dire à l'instant, sur cette question de l'extinction. Une hypothèse psychologique plutôt qu'institutionnelle sur ce qui peut motiver des Altman et compagnie à annoncer l'extinction. C'est aussi se donner à eux-mêmes une importance démesurée. Ils sont, au fond, ceux qui, par leur génie entrepreneurial, technologique et scientifique, sont en mesure, en quelque sorte, de provoquer l'extinction. Alors évidemment, ils disent : « mais ce n'est pas ce qu'on veut faire, mais voyez, on en est capables. Donc, on est vraiment les gens les plus puissants du XXIe siècle, on bat les Oppenheimer, etc. » Donc, je pense qu'il y a aussi une dimension de « se monter le bourrichon », si je puis m'exprimer ainsi, avec cette idée d'extinction. Qu'est-ce que tu penses de cette idée psychosociale ?

Raja Chatila [21.11]

Alors effectivement, ces entrepreneurs sont connus pour leur hubris, et effectivement, là on atteint un niveau assez important dans une croyance qu'ils peuvent avoir sur leur toute-puissance. Tu as raison de considérer cette dimension, parce qu'elle est importante dans leur comportement. C'est-à-dire qu'ils ont certainement l'impression de pouvoir influencer l'humanité (ce n'est pas complètement faux, d'une certaine façon) d'une manière tellement profonde et globale que, oui, psychologiquement, comme tu dis, ça leur est monté un peu à la tête et ils croient peut-être effectivement qu'ils ont des capacités importantes, y compris celle de l'extinction de l'humanité, se mettant au même niveau que les concepteurs de la bombe atomique. Je pense que c'est un élément qui en plus rejoint probablement une idéologie transhumaniste sous-jacente qui pénètre beaucoup ces grands entrepreneurs de la Silicon Valley. Oui.

Comprenons-nous bien ces systèmes ?

Daniel Andler [22.44]

Et si je peux, avant de rendre la parole à Mehdi, pour revenir au début de ton exposé, qui était tout à fait pertinent et particulièrement clair, je te remercie, tu as présenté les modèles génératifs de manière très sobre, en quelque sorte, en disant que nous avons fait pour le langage ce que nous avions fait auparavant pour les images, pour les raisons que tu as évoquées. C'était difficile, mais nous avons

surmonté les difficultés techniques et maintenant nous disposons d'un instrument qui produit un effet massif. Comme tu l'as très bien dit, les gens en général trouvent que quelque chose qui converse, qui parle avec pertinence, ressemble beaucoup à un être humain, avec son intelligence, et c'est très impressionnant. Je crois que tu as mentionné à juste titre que c'est très impressionnant non seulement pour le grand public, mais aussi pour nous, les chercheurs, les philosophes, les technologues, les scientifiques qui étudions ces systèmes.

Cependant, il y a une chose que tu n'as pas dite et sur laquelle je voudrais t'interroger. Il m'a semblé que tu as donné l'idée qu'au fond on comprenait bien ces systèmes. Or, nous ne les comprenons pas si bien que ça. Je voulais t'interroger là-dessus et aussi te poser une question qui me tourmente personnellement. Je n'arrive pas à comprendre comment ChatGPT arrive à distinguer l'objet et les méta-instructions. Autrement dit, comment est-ce qu'il parvient à comprendre que lorsqu'on lui dit « fais-nous un exposé de la relativité générale pour des personnes qui ne comprennent pas les mathématiques » ou « écris un sonnet érotique à la manière de la Déclaration des droits de l'homme », il doit distinguer que l'objet est le sonnet érotique ou la relativité générale, et qu'il doit tenir un discours sur cet objet, mais le faire dans un certain style, d'une certaine façon. Il arrive d'une certaine façon à ségréger les deux types d'instructions. Je n'arrive pas à comprendre comment il parvient à faire cela.

Raja Chatila [25.18]

Alors, évidemment je ne sais pas tout, loin de là, sur le fonctionnement de ChatGPT, et je ne peux pas répondre de manière bien argumentée sur la façon dont il analyse les phrases, qui peuvent être longues et complexes, pour répondre. Mais c'est aussi de l'ordre de quelque chose qu'on sait faire, qui est l'analyse des phrases, avec une structure sujet-verbe-complément, etc., et le parsing des phrases doit avoir des instruments. C'est une interface finalement. C'est l'interface de l'agent conversationnel, quand on interagit avec lui. Cela ne nécessite pas encore toute la puissance du modèle Transformer qui est derrière pour interpréter la phrase, le sujet, de quoi l'on parle, etc. Ensuite, et là je ne sais pas exactement à quel moment, assez rapidement après l'analyse de la phrase, il y a effectivement l'appel au modèle et donc la puissance des corrélations qui vont être faites pour produire la réponse. Je n'ai pas vu de publications spécifiques à ce sujet, mais je pense qu'il y a d'abord cette interface, qui peut être relativement classique, même si elle est puissante, d'interprétation des phrases, des prompts, et tout de suite augmentée par la capacité corrélative et la génération ensuite.

Daniel Andler [27.34]

D'accord. Merci, merci. D'ailleurs, tu as mentionné que tu n'avais pas tout lu, ce qui est compréhensible, car il est impossible de tout lire étant donné l'immensité de la littérature. Cependant, il faut également souligner que certaines parties de cette littérature sont dissimulées en raison du secret industriel, ce qui peut nous mettre mal à l'aise en tant que chercheurs.

Raja Chatila [27.40]

Oui. Heureusement, d'une certaine façon, il n'y a pas que ChatGPT, il y a aussi d'autres systèmes génératifs, qui adoptent une approche de logiciel ouvert et sont donc un peu plus transparents, même si les concepteurs ne divulguent pas tous les mécanismes sous-jacents. Aujourd'hui, on parle beaucoup de ChatGPT en raison de sa puissance et du choc initial qu'il a créé, ce qui lui a valu la première

place. Cependant, il existe effectivement d'autres systèmes qui arrivent sur le marché ou qui sont déjà là.

En plus, cela me donne l'occasion de parler des questions liées aux langues utilisées, car ces systèmes sont principalement basés sur l'utilisation d'un corpus en anglais. Il y a de nombreux projets visant à développer des systèmes génératifs dans différentes langues, comme le français, d'autres langues européennes, etc.

D'ailleurs, aujourd'hui (ou plutôt demain), lors du salon Vivatech, le président de la République devrait s'exprimer et aborder ces questions, notamment celles liées à l'IA générative, mais aussi des questions de souveraineté. Ceci est un concept qui englobe de nombreux aspects, en particulier la défense de la culture et de la langue française. Nous assisterons certainement à des développements de systèmes, disons, encouragés par différents gouvernements pour défendre leur langue et se faire une place dans la diffusion de ces technologies. On parle peu de la Chine dans ce contexte, car la Chine a aussi développé des systèmes, principalement en mandarin. Cependant, bien sûr ces systèmes ne sont pas diffusés en Occident. Cependant, et c'est un point essentiel, les corpus étant surtout en anglais, ils véhiculent la culture anglo-saxonne. Même s'il y a un pourcentage beaucoup plus faible de textes dans d'autres langues comme le français, l'allemand, etc., il s'agit d'un corpus plus restreint et choisi, donc il n'est pas nécessairement représentatif de toute la culture de ces langues. Parfois, l'anglais est également utilisé comme langue pivot, c'est-à-dire que l'on traduit vers l'anglais, puis on retraduit dans une autre langue, ce qui fait que, par ce mécanisme (c'est pour ça que je parle de souveraineté) on peut avoir des systèmes qui vont diffuser encore plus la culture et la manière de penser (disons, puisque c'est exprimé dans les textes) anglo-saxonnes. C'est un enjeu important.

Illusion d'intelligence des systèmes conversationnels

Mehdi Khamassi [30.51]

J'aimerais vraiment revenir sur quelque chose que tu as soulevé, à savoir l'illusion d'intelligence que peuvent susciter ces systèmes conversationnels. Ils donnent l'impression de comprendre les sujets dont ils parlent, ce qui peut poser des risques en induisant une confusion chez les utilisateurs. Cette illusion de compréhension peut rendre plus difficile la distinction entre ce qui peut être vrai et ce qui peut être faux dans ces systèmes.

Il me semble que tu as mentionné un point lors de nos discussions au sein du laboratoire qui mérite d'être partagé ici, à savoir l'incapacité de ces systèmes à comprendre et à raisonner sur les implications des énoncés qui manipulent des événements dans le monde physique. Par exemple, un énoncé qui parle du déplacement d'un objet dans l'espace et des conséquences physiques que cela entraîne semble être hors de portée de ces systèmes. Pourrais-tu en dire plus à ce sujet, s'il te plaît ?

Raja Chatila [31.47]

Oui, c'est un point essentiel, il me semble, car j'affirme, et on le sait, ces systèmes ne comprennent pas ce qu'ils disent. Ce problème de la sémantique, de la signification, est important en intelligence artificielle depuis toujours. Mais là, avec l'impression qu'on a quelque chose de construit dans la langue, on a parfois l'illusion que le système comprend de quoi il parle. Mais non, il ne peut pas réellement comprendre de quoi il parle. C'est inhérent à la nature du système, parce que ce sont des sys-

tèmes corrélatifs, et non pas causaux, capables de corréler des éléments, de décrire et de reconnaître, par exemple sur une image un objet particulier, sans que le système sache réellement ce que c'est.

Je prends l'exemple classique du traitement d'images avec un chat. Le système peut reconnaître des chats dans toutes les directions après avoir été entraîné sur une grande quantité de données étiquetées comme étant des chats. Cependant, cela ne suffit pas, car le système ne sait toujours pas ce qu'est un chat. Il ne possède jamais l'expérience phénoménologique du chat. Contrairement aux êtres humains qui vivent dans le monde réel, dans le monde physique, qui sont soumis aux contraintes du monde physique et qui interagissent avec lui, et dont les concepts les plus abstraits sont nés, ont été élaborés à partir d'une connaissance du monde physique.

Peut-être que Daniel, en tant que philosophe, ne sera pas d'accord, mais je pense tout de même qu'un cerveau dans un vase, même s'il est connecté uniquement à des yeux ou des caméras, aura beaucoup de mal à élaborer des concepts du monde physique qu'il observe. En tout cas je ne vois pas comment il pourrait y arriver. L'histoire des sciences s'appuie sur l'expérimentation, l'hypothèse scientifique, et pas seulement sur l'observation pure et simple. Toute cette démarche d'acquisition de connaissances, et d'abstractions des connaissances en une théorie, qui a un pouvoir explicatif et prédictif, s'est construite justement grâce à l'interaction avec le monde. Et à partir de là, on peut élaborer des concepts aussi complexes et aussi peu clairs que la mécanique quantique ou la dignité humaine, qui ne sont pas directement observables dans le monde, mais peuvent être construits, élaborés.

Donc, cette question de sémantique est essentielle, et pour moi l'absence de sémantique est inhérente à ces systèmes. En d'autres termes, quand ChatGPT dit « chat », il ne sait pas de quoi il parle. Cela pose un problème, car en ne sachant pas de quoi il parle dans le monde réel, le système peut dire des choses, produire des informations ou des textes issus de ces corrélations qui ne sont pas seulement inexactes, mais qui peuvent également amener les êtres humains à agir de manière inappropriée, voire dangereuse.

Entre parenthèses, c'est peut-être de là que vient l'histoire de l'extinction, c'est-à-dire que les systèmes peuvent convaincre les humains de s'éteindre d'une certaine façon. Mais ce qui commence à m'interroger sur ce que je viens de dire, en fait, c'est que des gens comme Geoffrey Hinton (de nouveau l'un des grands acteurs des systèmes d'intelligence artificielle modernes basés sur l'apprentissage, etc.) disent que ce n'est pas si clair que cela que ces systèmes ne peuvent pas élaborer une sémantique. Ce n'est pas forcément exactement la même que la nôtre, mais le fait que ce soient des systèmes purement corrélatifs ne signifie pas nécessairement qu'ils ne peuvent pas avoir une sémantique définie. Après tout, et là c'est moi qui le dis, pas lui, nous [humains] élaborons des théories, mais c'est bien parce que nous observons des corrélations que nous essayons de creuser davantage pour les comprendre. Ainsi, la corrélation en soi peut être porteuse d'un certain sens, d'une certaine sémantique. Je suis à moitié convaincu, enfin, plutôt beaucoup moins que la moitié, car on peut observer des pommes tomber toute la journée, élaborer la théorie de la gravitation c'est une autre affaire, et cela fait appel à ce pouvoir d'abstraction qui est réellement absent dans ces systèmes. Néanmoins, je crois qu'il faut peut-être creuser un peu plus cette question du passage du corrélatif au sémantique.

Et donc je reviens un peu sur le fondement de la sémantique, parce que la sémantique dans les langues, ce sont des concepts complexes. Tout le monde n'est pas d'accord sur la manière dont nous [humains] élaborons la signification des mots et du texte. Certains linguistes affirment que le contexte d'un texte, c'est-à-dire les mots qui se trouvent dans le même texte ou dans son voisinage, suffit à définir la sémantique d'un mot. Par exemple, quand je parle de mon chat, j'utilise des verbes, des adjectifs, des qualificatifs, et ce sont ces mots qui entourent le mot « chat » qui vont déterminer sa signification dans le cadre de mon discours. Et si j'ai plusieurs discours différents sur le chat, peut-être que, globalement, ce réseau de discours différents sur le chat est la sémantique du chat.

Daniel Andler [39.19]

Si je peux me permettre, cela m'intéresse beaucoup, car dans mon bouquin (ce n'est pas de mon bouquin qu'on parle, mais ça m'intéresse beaucoup), j'ai développé un peu cette idée. Justement, ce que j'appelle la « cécité sémantique » semble être similaire à ce que tu dis, à savoir que les systèmes ne savent pas vraiment qu'il s'agit d'un chat, mais plutôt qu'ils le reconnaissent comme une étiquette. J'appelle cela la « cécité sémantique » et je dis que finalement ce n'est pas aussi clair que ça, effectivement exactement selon ce que tu viens de développer à l'instant. Cela m'intéresse de savoir que Hinton lui-même a commencé à réfléchir à cela, et comment nous-mêmes parvenons à comprendre certains aspects de la sémantique. C'est vraiment très intéressant, et je crois que beaucoup de choses vont se développer, j'espère, sur cette question qui est assez obscure et profonde. C'est particulièrement intéressant pour les philosophes, je dois le dire, mais pas seulement.

Bientôt un remplacement de toutes sortes de métiers ?

Daniel Andler [40.17]

Je profite du fait d'avoir la parole, parce qu'on a très peu de temps, pour poser la question suivante : nous savons donc que ces systèmes ne sont pas fiables, qu'ils peuvent avoir ce qu'on appelle « des hallucinations », c'est-à-dire que par moments ils inventent des trucs ; au milieu de quelque chose de parfaitement correct, tout à coup ils sortent quelque chose qui n'existe pas, qui est mystique, qui est complètement erroné ou complètement déplacé. Crois-tu que ces systèmes pourront peu à peu s'améliorer au point de devenir fiables ? Si c'est le cas, cela aura un impact sur les médecins, les juristes et toutes sortes de métiers ; ils pourront faire faire par un super ChatGPT ce qu'ils font actuellement avec leur intelligence humaine. Est-ce que tu vois une limite de principe ? Qu'est-ce que tu en penses ?

Raja Chatila [41.07]

Je crois qu'il y a une limite de principe, et je crois qu'il y a en même temps moyen de repousser cette limite, effectivement. La limite de principe est toujours la même : c'est l'aspect corrélatif. Aujourd'hui, sans même parler de ChatGPT, etc., mais on y reviendra, tous ces systèmes qui reposent sur l'analyse de données pour établir des corrélations entre les éléments constitutifs de ces données, les classer, générer de nouveaux éléments, etc., tous sont basés sur de la statistique. Certes, une statistique très élaborée, une grosse fonction mathématique qui est le réseau de neurones, qui élabore ce modèle statistique. Ça n'en reste pas moins une statistique. Cela signifie que ces systèmes construisent un modèle qui est un espace dans lequel il existe des vecteurs représentant les données qui ont été analysées et classées dans certaines zones de cet espace. Le système interprète ensuite ces données en mesurant une distance entre deux éléments dans cet espace de très grande dimension. Si la distance est faible, cela signifie que ces objets sont voisins, voire confondus. Or, cette proximité, qui est fondée sur une corrélation, n'est pas toujours appropriée ou réelle dans la réalité. Nous pouvons voir de multiples exemples où une image d'un scooter sur une route avec un ciel bleu et une prairie à l'arrière est interprétée avec une certitude (c'est-à-dire un degré de confiance, plus exactement) de 99% comme étant un scooter. Et on change légèrement l'attitude

du scooter et l'interprétation devient un parachute avec un degré de confiance de 100%. On voit bien que là il y a quelque chose qui s'est passé dans cet espace, où la confusion n'était pas due à l'observation du scooter, mais peut-être du paysage qui est derrière. Cette proximité a joué. Ainsi, il y a un problème inhérent à ces systèmes : on peut tout faire, mais dès lors que c'est statistique, et même avec des pourcentages aussi élevés que 99%, les systèmes peuvent se tromper, se tromper de manière non prévisible, et se tromper avec aplomb, en disant « je suis sûr, c'est 1 ! ». Et ça, ça va rester. Cependant, il y a une possibilité de progrès grâce à l'apprentissage continu. Si nous injectons une correction dans le système suite à une corrélation erronée, par exemple en disant que ce n'est pas un parachute mais un scooter (ce que je dis est bien sûr imagé), cela signifie que nous pouvons modifier les distances dans cet espace de façon à améliorer la classification. Bien sûr, il s'agit de systèmes hautement non linéaires, donc la perfection ne sera jamais atteinte, mais on peut peut-être faire des ajustements. À long terme, je pense que si on adopte ce type d'approches, ce qu'on appelle l'apprentissage continu, qui n'est pas trivial, alors on peut imaginer que ces systèmes vont s'améliorer.

Alors déjà dans ChatGPT il y a quelque chose qui a été introduit : vous lui posez une question au lieu de demander de faire quelque chose et il produit un résultat faux. Par exemple, dans un raisonnement logique ou un calcul, il peut donner une réponse erronée. Eh bien, l'ingénierie du prompt, consiste à dire maintenant « non, ça, c'est faux », et de procéder pas à pas, « step by step » en anglais. Donc on va lui donner finalement les démarches, en d'autres termes on va guider les corrélations qu'il faut adopter pour arriver à la solution. Lorsqu'on lui donne ces indications, il donne de bonnes réponses.

Daniel Andler [46.21]

Passionnant !

Raja Chatila [46.23]

Cependant, je m'empresse de dire qu'il ne faut surtout pas comparer cela à la manière dont on explique des choses à un autre être humain ou à un enfant. Ce n'est pas du tout comme ça qu'il faut voir les choses. Mais il y a quand même l'idée que si nous pouvons guider le traitement que le système effectue dans le but de l'améliorer, alors il va s'améliorer. L'approximation sera meilleure jusqu'à un certain point, peut-être, où cela deviendra tellement meilleur, c'est-à-dire que ça ne sera pas 99 % mais avec plusieurs 9 derrière, que cela pourrait être quasiment acceptable. Acceptable en termes de résultats, car de toute façon, notre connaissance de la réalité n'est jamais parfaite, et nous pouvons l'admettre, car tout dépend aussi de la manière dont nous percevons et acceptons les choses. Je sais très bien que si je monte dans un avion il y a un risque qu'il s'écrase, mais ce risque est tellement mineur que je le prends. Donc nous pouvons croire beaucoup de choses, et dans le domaine de la médecine, puisque la médecine a été évoquée, ce n'est pas une science exacte, il y a toujours des risques, on le sait bien. Mais si les risques sont vraiment minorés, on pourra peut-être prendre ces risques. Ce qui veut dire que ces systèmes ne sont pas inutiles, ils ne doivent pas être rejetés, comme ça, d'un coup en disant « non, non, ça ne va pas, ça ne va pas ». Il faut au contraire soulever leurs limites pour justement les améliorer et les utiliser surtout, surtout à bon escient.

Juste pour conclure sur la médecine, et cela est lié à notre conversation, il y a un article récent dans le New York Times qui dit que certains médecins ont utilisé ChatGPT, mais qui ont bien vu que côté médical, à proprement parler, ça n'était pas extraordinaire. En revanche, cela a amélioré leur capacité

à communiquer de manière empathique avec les patients. En d'autres termes, on a appris l'empathie d'une machine. Je trouve ça à la fois absolument (rires) absolument choquant, finalement, sur ce que ça révèle de l'état de l'empathie, et en même temps, cela pose des questions.

Daniel Andler [48.52]

Merci.

Risques réels, dont la fracturation de la société

Mehdi Khamassi [48.58]

Nous avons encore quelques minutes pour conclure sur les risques réels. Si on ne reste pas que du côté des risques existentiels, qui peuvent susciter des réactions émotionnelles voire même, comme tu l'as dit, nous empêcher d'envisager sereinement les problèmes de gouvernance pour légiférer, on peut se poser la question de quels sont les risques réels, qu'il faut vraiment prendre en main, et quelles régulations il faut envisager.

Daniel Dennett écrivait récemment qu'un des problèmes majeurs pour les sociétés humaines est la contrefaçon ; contrefaire un discours, se faire passer pour un humain, ce qui peut saper la confiance et les fondements de la société aussi fortement que pouvait l'être la contrefaçon de la monnaie. Naomi Klein, dans un autre texte récent, souligne le problème du droit d'auteur quand on entraîne ces systèmes sur des œuvres générées par d'autres, des êtres humains, pour ensuite pouvoir générer de nouveaux contenus qui leur ressemblent, à la manière de, pouvoir faire des profils là-dessus. Dans le même temps, les personnes qui ont créé les œuvres initiales et ont contribué à leur entraînement ne sont pas rémunérées, voire même leurs œuvres deviennent moins visibles. Est-ce que, pour toi, cela fait partie des risques importants sur lesquels il faut légiférer ? Et quels autres risques tu vois ?

Raja Chatila [50.01]

Le risque de la contrefaçon, c'est le risque de la vérité. Je l'interprète comme ça, un peu. La contrefaçon, c'est imiter quelque chose qui existe par ailleurs et qui est vrai, à l'aide de quelque chose qui est faux. Je ne sais pas dans quel sens Daniel Dennett l'a utilisé, mais c'est peut-être un peu limitatif par rapport au problème de l'invention, comme vient de le dire Daniel sur les hallucinations. Il s'agit plutôt de l'invention de la vérité. Ce n'est pas de la contrefaçon, c'est du faux, du faux inventé, mais pas de l'imitation du vrai. C'est quelque chose de plausible néanmoins, mais complètement différent. Ce n'est pas simplement répliquer un sac en cuir d'une marque célèbre avec une matière peu noble, c'est vraiment quelque chose qui a été complètement inventé. Je ne sais pas si le terme de « contrefaçon » est tout à fait approprié dans ce contexte. Mais oui, je pense que ce risque est profond pour nos sociétés.

Une société donnée, et donc un pays en général, voire un ensemble de pays, une civilisation d'une certaine façon, est fondée sur un certain nombre de croyances, de concepts, de faits réels historiques ou quotidiens, de suppositions qui sont partagées. Elles ne sont pas listées quelque part, mais elles sont partagées. Ce socle commun est un liant de la société, car il nous permet d'être d'accord sur un minimum, sur ce socle, et à partir de là, d'avancer ensemble. Si ce socle est fracturé, remis en

question, fracturé parce que chacun a le sien, finalement, étant donné qu'il a interagi avec un système génératif et et que le système lui a répondu quelque chose dans lequel il a confiance, dans lequel il croit, mais qui n'est pas avéré dans la réalité, et en plus, le système n'a pas répondu la même réponse à l'un ou à l'autre, alors notre socle est fracturé. C'est comme la banquise qui commence à se fracturer, on ne peut plus établir un pont, un lien, une possibilité d'être ensemble, d'être d'accord sur un minimum. Et cette fracture peut fracturer la société dans son ensemble. Donc, là il y a un danger réel. Une extinction ? Je ne crois pas. Mais un réel danger. À mon avis, on a vu les éléments ou les prémices dans l'assaut du Capitole. Ce n'était pas de l'IA générative, mais plutôt des réseaux sociaux, qui ont provoqué le fait que des gens croient dur comme fer quelque chose que quelqu'un a dit, mais qui s'est propagé, qui a constitué pour eux cette expression oxymorique de la « vérité alternative ». À partir de ce moment-là, oui, il y a un danger réel pour nos sociétés. Déjà qu'il y a très peu de confiance, par exemple, dans les personnalités politiques. Cela peut encore s'aggraver, et cela ne sera plus seulement une absence de confiance dans les personnalités, ce sera une absence de connaissances communes et de confiance dans les faits, dans la réalité des choses, de ce qui s'est passé, de ce qui se passe, des événements. Et cela peut être extrêmement dangereux. C'est l'un des sujets les plus importants, je crois, la question de la vérité. Nous avons parlé de la sémantique, mais nous avons aussi évoqué cette question de la vérité. Effectivement, sur ce plan je suis d'accord avec Denett ; c'est une menace assez majeure. C'est pourquoi il est nécessaire de gouverner, de savoir gouverner et de réglementer ces systèmes, bien que ce ne soit pas très clair comment le faire concrètement.

Mehdi Khamassi [54.19]

Merci beaucoup Raja, je pense que ce sera le mot de la fin, puisqu'il est temps de retourner à nos occupations. En tout cas, voilà de quoi contribuer à faire réfléchir et à entretenir ces débats. Donc à très bientôt !

Raja Chatila [54.30]

Merci à vous !

Daniel Andler [54.32]

Merci beaucoup, Raja. C'était passionnant !

L'IA est-elle soutenable ?

Audition de Michèle Sebag

MICHÈLE SEBAG

Michèle Sebag est Directrice de Recherche (DR) au CNRS. Elle co-dirige l'équipe TAU (Tackling the Underspecified) de l'INRIA avec Marc Schoenauer, sur le campus de l'Université Paris-Saclay. Elle a une formation initiale en mathématiques à l'École Normale Supérieure, puis une expérience d'ingénieure et une thèse en informatique. Elle est membre de l'Académie des Technologies depuis 2017. Spécialiste de l'apprentissage automatique (dit apprentissage machine, machine learning en anglais), elle enseigne cette discipline depuis des années en master d'informatique à l'Université Paris-Saclay. Elle a publié de nombreux articles scientifiques, édité plusieurs ouvrages et écrit de nombreux chapitres de livres. En 2021, elle a contribué et fait une synthèse du colloque interdisciplinaire LINX de l'École Polytechnique (organisé par François Levin et Étienne Ollion) sur les enjeux et ce qui échappe à l'Intelligence Artificielle (IA), dont les vidéos sont accessibles en ligne : https://www.youtube.com/watch?v=g3xvpHVAkE8.

L'audition a été menée par Mehdi Khamassi et Daniel Andler

1. Sur les grands modèles de langage

Mehdi Khamassi [0.08]

Merci beaucoup, Michèle Sebag, d'avoir accepté cette audition pour TESaCo, le projet Technologies Émergentes et Sagesse Collective, qui est dirigé par Daniel Andler, ici présent, pour l'Académie des Sciences Morales et Politiques. Je voulais, avant qu'on commence à te poser des questions, faire une brève introduction : dire que tu es directrice de recherche au CNRS, que tu rediriges une équipe INRIA avec Marc Schoenauer sur le campus de l'Université Paris-Saclay. C'est une équipe qui s'appelait anciennement TAO, et qui est aujourd'hui nommée TAU, pour Tackling The Unspecified, donc aborder des questions non spécifiées, si je traduis bien, et si tu es d'accord avec cette traduction.

Michèle Sebag [0.45]

En fait, c'est Underspecified. Mais c'est exactement l'idée.

Mehdi Khamassi [0.47]

Ah ! D'accord ! Et puis tu as une formation d'abord en mathématiques à l'École Normale Supérieure, et puis une expérience d'ingénieure, et une thèse en informatique. Tu es membre de l'Académie des Technologies depuis 2017. Et ce qui est très important pour nous dans le cadre de cette audition sur l'IA, sur l'Intelligence Artificielle, tu es spécialiste d'apprentissage machine, qui est quelque chose sur lequel tu as beaucoup publié, beaucoup travaillé, et que tu enseignes aussi depuis des années, notamment au Master d'informatique à l'université Paris-Saclay. Et… (Interruption)

Michèle Sebag [1.20]

Pardon ! Tout ce que tu as dit est vrai.

Mehdi Khamassi [1.22]

Super ! (rires) Je suis content de pas m'être trompé. Et donc on avait envie de discuter avec toi aujourd'hui des progrès de l'IA, des progrès fulgurants de ces dernières années. Notamment, il y a eu des progrès techniques, mais qui se sont accumulés sur beaucoup plus d'années que l'apparence, que la visibilité très forte de l'IA depuis à peu près une dizaine d'années, notamment en dehors du milieu académique. On voulait discuter à la fois d'aspects techniques, peut-être d'écart entre ce qu'on en comprend à l'heure actuelle, ce qu'on comprend du fonctionnement, de ce que ça permet de réaliser, et puis de comment ça peut interagir avec la société. Et nous voulions aussi te poser des questions – là, je fais vraiment juste une vue d'ensemble –, des questions d'applications possibles pour la société, de soutenabilité de ce type d'approches ou de technologies, et des questions d'éthique, bien sûr.

Écart entre les réalisations de l'IA et ce qu'on en comprend

Mehdi Khamassi [2.15]

J'aurais voulu commencer avec toi sur ce constat de réalisations extraordinaires de l'IA, et en même temps d'un écart qui semble constaté par pas mal de monde entre ce qu'on arrive à faire avec l'IA et ce qu'on comprend de pourquoi ça fonctionne. Dans quelle mesure cet écart empêche en partie de bien anticiper comme l'IA interagit avec la société ? Est-ce que tu partages ce constat ?

Michèle Sebag [2.41]

Oui, mais je pense que ceci est assez ancien. Ce n'est pas à toi que je parlerai d'Eliza, le premier programme d'IA de [Joseph] Weizenbaum [dans le milieu des années 60], qui, si ma mémoire est bonne, avec des trucs extrêmement simples, réussissait à faire croire à son interlocuteur qu'il était compris. Et je pense que c'est une de nos spécificités [en tant qu'humains] : supposer qu'il y a du sens, où qu'on soit. Et donc, tant que l'IA n'aura pas fait la preuve qu'il s'agit grosso modo d'un singe et non d'une vraie intelligence artificielle, je crois que les gens y croiront. Et je pense qu'il est difficile

aujourd'hui de ne pas parler du ChatGPT, qui se présente comme un super Google dans le sens où non seulement il donne les informations, mais il en fait un résumé audible assez peu conflictuel, ce qui est sa mission. Donc il est difficile de ne pas lui prêter de l'intelligence.

Ce qui est paradoxal, c'est qu'un bon nombre d'amis ont essayé de pousser ChatGPT – le modèle de langage qui vient de sortir il y a, je ne sais plus, avant Noël –, ils ont essayé de le pousser dans ses retranchements pour voir où était la frontière de son savoir. Ce qui est très intéressant, et à mon avis c'est une distinction fondamentale par rapport aux produits technologiques dont on a l'habitude, c'est que tester ChatGPT, c'est directement contribuer à le renforcer, c'est-à-dire à combler ses trous. Et si on imagine que X centaines de milliers de personnes sont en train de le tester et de voir où il se trompe et comment il discute, il s'agit à mon avis à ce moment-là de l'entreprise de test la plus grande que l'humanité a connue.

Donc là nous avons un objet nouveau, qui parle, et qui répond littéralement à des centaines de milliers de gens qui lui posent des questions, essayant de percer jusqu'où il comprend. Autant on peut dire quelque chose sur ses trous à l'heure actuelle, autant c'est difficile de dire ce que pourront être ces trous quand ils auront été patchés justement à l'aide des interactions.

Tester ChatGPT c'est contribuer à l'améliorer

Daniel Andler [6.07]

Michèle, puis-je te poser une question technique, un peu élémentaire, mais peut-être pas si élémentaire que ça ? J'ai à peu près compris comment fonctionnent les Large Language Models (LLMs ; en français, les grands modèles de langage). Je regarde depuis un moment, puisque ChatGPT est juste le dernier. Je comprends bien comment on constitue la base d'apprentissage initial : on regarde tout ce qu'on peut trouver sur internet, Wikipédia, etc. On fait tourner la machine. C'est un processus techniquement assez complexe mais néanmoins on voit à peu près comment ça marche. Ce que je ne comprends pas encore bien, moi qui ne suis pas informaticien, c'est comment justement s'effectue cette amélioration en cours de fonctionnement, c'est-à-dire pas en cours de l'entraînement initial, mais au cours du fonctionnement. Donc exactement ce dont tu viens de parler : à savoir qu'à mesure que ChatGPT est challenged – i.e., poussé dans ses retranchements par les utilisateurs – il s'améliore. Comment s'effectue en gros cet apprentissage au cours de l'utilisation avec les utilisateurs ?

Michèle Sebag [7.24]

Écoute, je ne suis pas dans le secret de ChatGPT. Mais voici la manière dont ça se produit pour des systèmes comparables. Tu as des labelers, donc le rôle du Turk d'Amazon : le fait de faire appel à des gens qui étiquettent ce qui se passe. Ceci est un travail massif assez ignoré derrière tout ce qui est [visible]. Le même travail peut être appliqué du côté de la production de phrase. Donc tu pourrais avoir des gens qui étiquettent des phrases comme étant acceptables, bonnes, ou bien limites. Et ceci pourrait produire le même effet social par rapport à la production de phrases de la machine, qui indique ce qui va être bien accepté ou non de l'auditoire. Dans le cas des erreurs de ChatGPT, d'abord, si tu veux mon avis, ils ont intérêt à garder quelques erreurs ; le fait de montrer que ce système a malgré tout des pieds d'argile peut être quelque chose de rassurant pour l'humanité, malgré tout. Pour prendre un exemple de la façon dont on peut le coincer, et donc dont on pourrait réparer ça : si on pose

la question à ChatGPT « est-ce que Ab[raham] Lincoln a bu de l'eau ? », la réponse est « oui ». Si on pose la question « est-ce que Lincoln a mangé du pain ? », la réponse est « je ne sais pas » [alors qu'il est évident que Lincoln a déjà mangé du pain dans sa vie, comme il a aussi bu de l'eau]. Que ferions-nous si nous étions les concepteurs de ChatGPT ? Indiquer que l'eau et le pain sont au même niveau ? Ce n'est pas tout à fait exact, car il y a des continents entiers où il n'y a pas de pain. Tu vois ? Donc qu'est-ce que tu vas rectifier là ? Si quelqu'un est dans une certaine partie du monde, alors le fait de manger du pain est à peu près aussi probable que de boire de l'eau. De cette manière-là tu peux rectifier les contours de ce que sait, croit savoir, et de ce que dit en particulier ChatGPT.

Daniel Andler [10.13]

D'accord. Oui, je comprends. Merci. Peut-être que j'enchaîne. Je ne sais pas combien de temps il faut qu'on passe sur ChatGPT, parce qu'il y a beaucoup de choses à dire. Mais une des choses qui me frappent – pas seulement pour ChatGPT, mais déjà pour d'autres systèmes que j'ai regardés un peu –, ce n'est pas tellement que sur un sujet donné il arrive à faire une dissertation à peu près au niveau d'un étudiant moyen qui aurait une mention assez bien, une mention bien de chez nous, mais c'est une sorte d'à propos. Je donne un exemple : mon père a demandé à ChatGPT « quelle est la différence entre la philosophie analytique et la philosophie continentale ? », qui est un geste standard. Et il a sorti une petite dissert pas mal du tout. Et puis mon père a dit : « bon, merci beaucoup, mais quand même, tu n'as pas du tout parlé de Wittgenstein. » Et là, ChatGPT a dit la chose suivante : « ça, c'est vrai, je me suis trompé. Je m'excuse beaucoup, j'aurais dû parler de Wittgenstein. » Et puis il sort une petite dissert tout à fait correcte mais pas sublime sur Wittgenstein. Et moi, disons, le semblant d'intelligence que je vois là n'est pas tellement dans la capacité à faire un petit topo sur Wittgenstein, ou sur la philosophie analytique, mais c'est dans l'à propos, l'à propos de la réponse : il dit « ah ben oui, c'est vrai, j'aurais dû penser à Wittgenstein. Je m'excuse, je complète ma réponse précédente ». Ceci a l'air très intelligent, justement, et je me demande comment c'est possible.

Michèle Sebag [11.52]

Écoute, je ne sais pas ; et je n'ai pas les connaissances. Maintenant, tu sais, il est très possible que ça soit encore un des méfaits de la chambre chinoise : si tu patches un certain nombre de réflexes, tu arrives à donner une impression assez éduquée, tu vois. Et effectivement le fait de reconnaître ses erreurs, ou le fait de le reconnaître sous cette forme-là, est extrêmement habile.

Daniel Andler [12.33]

Oui oui oui, tu as raison, ça fait beaucoup penser à Eliza, oui c'est vrai, tout à fait. Il [ChatGPT] a pris l'habitude, quand on lui répond comme ça méchamment, dans une objection, à dire « oui c'est vrai, désolé, j'aurais dû y penser ». En tout cas, oui, merci, c'est une très bonne explication, plausible, car on ne sait pas exactement comment ça marche. Mais d'accord.

Mehdi Khamassi [12.54]

Et on est tenté de penser que statistiquement, peut-être que les humains qui ont contribué à tester la machine, et donc à l'entraîner, à constituer la base d'entraînement, avaient ce type de réponse quand

ils avaient oublié de compléter quelque chose. Donc c'est peut-être parce que c'est statistiquement très fréquent pour les humains de répondre ce type de réponse que la machine a fini par l'intégrer comme un réflexe, si je comprends bien ce que tu dis Michèle.

Michèle Sebag [13.15]

Oui, ça peut être ça, ou ça peut être le simple fait de dire [à la machine, donc à ChatGPT] que quand on vous dit quelque chose et que vous voyez que c'est pertinent, prendre une position comme celle que dit Daniel : « bon sang, mais c'est bien sûr, excusez-moi ! ».

Mehdi Khamassi [13.30]

Et en même temps s'il fallait faire ça pour toutes les situations possibles, ce serait infernal, parce qu'il y aurait des tonnes de situations où il y aurait des codes de réponses qui seraient acceptables. Donc ça veut dire que les concepteurs devraient avoir prévu à peu près toutes les situations à l'avance pour que ChatGPT réponde correctement.

Michèle Sebag [13.50]

C'est effectivement la question, et est-ce que c'est une infinité ? D'un côté, si tu interagis avec des centaines de milliers de gens, combien de temps te faut-il pour avoir un éventail de toutes ces situations ? Je pense que là on est dans les valeurs extrêmes ; je pense que le nombre de cas dans lesquels il n'a pas la parade va diminuer très vite. Et d'un autre côté, je pense bien entendu que les concepteurs du système savent que c'est un travail à plein temps ; il n'existe pas par construction de moments où on va pouvoir se dire « ça y est, c'est bon ». Grosso modo, ce à quoi je m'attends, c'est qu'il faille de temps en temps, comme chez le psychanalyste, que le système repasse par des entretiens avec ses concepteurs pour voir s'il y a des boucles, si on arrive à entraîner le système dans des boucles, disons, indésirables.

Les LLMs sont-ils des singes ?

Mehdi Khamassi [15.09]

Je suis frappé par un élément de la réponse que tu as donné au début, Michèle, sur le fait que tant qu'il y aura un écart entre ce que les gens perçoivent même de magique de l'IA et ce qui a été réellement conçu, tant que les gens ne comprendront pas que l'IA est simplement un singe, pour utiliser ta terminologie, ça posera problème. Et donc, même si on anticipe un peu sur les questions d'après, je trouve que c'est important à ce stade de réagir : qu'est-ce qui selon toi permettrait d'inciter les concepteurs à montrer de façon transparente que l'IA est simplement un singe, alors qu'il y a tous ces besoins de vendre, de démontrer que ça marche, et d'impressionner ?

Michèle Sebag [15.46]

Tu vois qu'actuellement le système a des phrases très humbles : « je ne suis qu'un lonesome système de conversation, etc. Je n'ai pas la vérité et ainsi de suite. » Je pense que tout cela est complète-

ment inopérant, tu vois. Ce sont des protections verbales qui permettent par exemple de dire « ben non, je ne vais pas vous donner la recette d'un cocktail Molotov. Qui suis-je ? Je ne suis qu'un malheureux système, etc., etc. ». Mais attends, je ne suis pas sûre d'avoir bien compris le sens de ta question. Qu'est-ce qui fait que les gens… Qu'est-ce que les concepteurs pourraient faire pour indiquer la différence entre ce singe et un être humain ? Je ne suis pas sûre qu'ils aient envie d'indiquer la différence.

Mehdi Khamassi [16.51]

Justement ! Comment pourrait-on inciter, encourager à le faire, si ça semble quelque chose de nécessaire pour l'intégration sociétale ?

Michèle Sebag [17.00]

Eh bien je n'ai pas beaucoup approfondi cette réflexion, mais je pense que tu vas justement m'aider à l'approfondir. Le système sur l'AI Act européen, si j'ai bien compris, correspond à une sorte de nutriscore de l'IA : voici des usages qui vont tout à fait bien ; voici des usages dans lesquels vous devez commencer à vous méfier ; et puis voici des usages qui sont carrément toxiques. Qu'est-ce que ça peut avoir comme effet ? Franchement, je ne sais pas. Il me semble qu'une partie des gens a toujours voulu un oracle, que nous sommes maintenant plus près que nous n'avons jamais été de cet oracle. Et je ne sais pas du tout ce qu'il faudrait pour que les gens renoncent à l'oracle. Ce n'est pas à toi que j'apprendrai ça, mais il y a pas mal d'études sur le rapport à l'incertitude, et le fait de pouvoir ou non l'accepter. Vivre dans un état d'incertitude, je crois comprendre, est éprouvant nerveusement, et une partie de notre économie psychique consiste justement à bâtir des certitudes pour réduire la fatigue [mentale]. Donc qu'est-ce qui peut convaincre les gens de leur plein gré de renoncer à un tel usage ? Franchement, je n'en sais rien.

Mehdi Khamassi [18.52]

Ce sont de grandes questions. Tu as mentionné l'AI Act, et donc là c'est la question de la régulation, qui est sûrement importante aussi. Et en plus il y a des questions d'échelles dans l'AI Act : dès que ce sont les très grandes entreprises, des algos ou des interfaces qui vont toucher un nombre d'utilisateurs considéré comme très grand, il faut être beaucoup plus transparent, il faut qu'il y ait davantage de règles. Maintenant, il y a aussi, et tu le soulèves, cette question de comment communiquer, encourager à ce que les gens, les concepteurs, considèrent eux-mêmes comme important de faire cette transparence. Dès lors, justement, dans quelle mesure cette transparence n'aiderait pas collectivement à réduire une part d'incertitude, et peut-être à faire qu'il y ait une intégration qui pourrait mieux se faire ? Je schématise, je simplifie peut-être. Peut-être que l'on se trompe à penser que c'est ça qu'il faut. Mais si on considérait que c'était ça qu'il fallait, est-ce qu'il y aurait à jouer sur ce terrain-là ? Peut-être dans la formation à l'éthique, à la réflexion sur les conséquences des sciences et des technologies qu'on peut développer, dans la réflexion sur l'apport des sciences à la société. Je ne sais pas. Qu'est-ce que tu en penses ?

Michèle Sebag [20.02]

Je pense que les concepteurs ont une vue qui est d'abord pragmatique. Je pense également qu'ils sont conscients des implications éthiques. Mais je ne sais pas dans quelle mesure le fun de faire

un système efficace ne l'emporte pas sur l'éthique. Tu vois, c'est quand même malgré tout un jeu extrêmement passionnant de faire un système. Faire un système éthique, avoir le câblage, je ne pense pas que ça soit faisable pour plusieurs raisons en particulier : le fait peut-être que l'éthique est quelque chose qui varie selon qu'on est en deçà ou au-delà de la montagne, en fonction des pays. Mais si tu considères que l'éthique est quelque chose qui se transmet socialement, alors le fait d'interagir avec des tas d'humains et d'avoir des instructeurs, qui épluchent ce qu'il faut regarder, ce à quoi il faut prêter attention dans les milliers de dialogues qui se déroulent, peut-être une façon de produire un effet. Mais qui je pense encore que ce sera un effet singe. Certains propos sont inappropriés, certaines lignes de raisonnement sont dangereuses, etc., etc. Donc évitons-les ! [D'une certaine façon], ça résout le problème, mais au fond on n'est pas sûr qu'on a compris le problème même si on l'a résolu.

Comment empêcher le moteur d'aller n'importe où ?

Mehdi Khamassi [22.09]

Pour rester sur une question d'éthique, dans les échanges qu'on a eu en préparation de cette audition tu as soulevé la problématique de comment empêcher le moteur d'aller n'importe où, et du coup d'interagir avec n'importe qui ? Et on peut même se demander : est-ce qu'il faut empêcher le moteur d'aller n'importe où ? Ou bien est-ce qu'on a besoin de le laisser aller là, justement pour apprendre de ses erreurs ? Et est-ce qu'il faut en même temps accompagner ça d'autre chose, de transparence, de précaution ? Qu'est-ce que tu crois qu'on doit faire de ce côté-là ?

Michèle Sebag [22.42]

Est-ce qu'il faut empêcher le moteur d'aller là ? Eh bien je pense que oui, c'est un jeu avec un adversaire. Tu vois par exemple, il est dit que le pire danger pour les véhicules autonomes est que les véhicules conduits par un humain meurent d'envie de taquiner, ou éventuellement plus méchamment, de mettre en difficulté le véhicule autonome. Donc ils lui font des tas de niches. En particulier, un des points qui est difficile consiste à arriver à s'insérer dans la ligne en sortant d'une place de parking, ou bien quand en arrivant sur l'autoroute. Dans ces cas-là notamment, les utilisateurs humains peuvent être assez durs pour tester la machine. Donc le fait qu'on doive mettre des barrières est indissociable du fait que l'on peut penser qu'il existe une partie des joueurs, des humains, qui ont envie de voir ce que le ChatGPT peut faire de pire. Je ne crois pas qu'on puisse aller contre. Je ne crois pas non plus qu'on puisse filtrer les interlocuteurs du système. Donc à partir du moment où il y a cette petite griffe du diable qui est « qu'est-ce que je vais pouvoir lui faire dire aujourd'hui ? », eh bien non, je ne pense pas qu'on puisse le laisser en liberté et évoluer tout seul.

Mehdi Khamassi [24.43]

Daniel, avant que je ne relance sur une question un peu plus scientifique ou technique, est-ce que tu avais envie de rebondir sur ces aspects, ou de poser une question dans ce domaine ?

Quelles sont les capacités actuelles des LLMs ?

Daniel Andler [24.53]

Je suis dans une position un petit peu curieuse vis-à-vis de ces Large Language Models, donc ChatGPT notamment. D'habitude je suis plutôt critique de l'IA et je dis : « L'IA, oui, bon, mais en fait c'est beaucoup moins. Voilà ! » Donc ma sympathie va de manière naturelle aux gens tels qu'Emily Bender, et tout ce qu'on a entendu jusqu'à présent dans l'interview : l'idée que c'est ChatGPT n'est pas vraiment intelligent, c'est une sorte de singe, etc. Mais néanmoins je continue de penser qu'il y a quelque chose de très mystérieux dans la manière dont un système qui a été entraîné sur un paradigme simplement de bouchage de trou, de complétion d'un mot manquant ou d'un fragment de phrase manquante, arrive malgré tout à produire des réponses intelligentes en combinant des « Fais-moi donc – je ne sais pas moi – un conte de fées dans le style de Shakespeare en introduisant des éléphants roses, ou je ne sais quoi ». Je continue d'être mystifié, en quelque sorte, par cette capacité. Et je ne suis pas entièrement sûr, en quelque sorte, qu'il suffise de dire que ChatGPT est une sorte de singe. Voilà. D'habitude, normalement, je devrais m'en contenter (rires) vue mon attitude déflationniste personnelle par rapport à l'IA. Mais j'avoue que je n'y arrive pas tout à fait dans le cas des modèles de langage. Mais cela dit, c'est juste pour pousser éventuellement à mon tour Michèle dans ses retranchements, car ce sont les idées de Michèle qui nous intéressent aujourd'hui.

Michèle Sebag [26.47]

Mais tu sais, le fait de dire qu'un singe est un singe n'empêche pas de considérer que les singes aussi sont intelligents. Le problème est de savoir quelle est la nuance entre les deux [l'intelligence du singe et celle de l'humain]. Il y a quelque chose qui a mystifié les humains depuis très longtemps : c'est la mémoire ! Tu te souviens, dans *Le Rouge et le Noir* [le roman de Stendhal], Julien Sorel a une mémoire eidétique et est capable de se sortir d'embarras en pêchant des fragments de Xavier de Maistre, ou je ne sais plus exactement qui étaient les auteurs de l'époque. Et ceci va lui donner une réputation d'intelligence énorme. Or ici, tu vois, l'intelligence du singe consiste à retrouver dans une mémoire énorme quels sont les fragments appropriés.

Daniel Andler [28.05]

C'est ça.

Michèle Sebag [28.07]

Donc j'essaie de voir quel est l'antipode de cette capacité d'esprit. Grosso modo, avec une excellente mémoire et le fait de savoir quel fragment est approprié dans quel contexte, je dirais qu'on peut s'en sortir dans 95% des cas en donnant l'impression d'être très intelligent. Non pas au sens du singe, mais au sens de l'humain. Qu'est-ce qui est aux antipodes, disons, de cet esprit, de ce système [, de ChatGPT] ? Tu pourrais dire que c'est le fait de faire des mots d'esprit. Par exemple, est-ce qu'il va pouvoir repêcher [un fragment] et adapter ? Mais aussi, est-ce qu'il va pouvoir plaisanter ? Est-ce qu'il va pouvoir faire de l'esprit au sens de, c'est notre vision probablement un peu idéalisée, au sens des cours monarchiques ? Mon impression est que c'est assez gris, que le système pourrait

aussi faire des traits. Je comprends tout à fait bien que personne ne s'y lance parce que je crois que le marché n'est pas fort grand. (rires) Mais je ne suis pas sûre qu'un singe ne pourrait pas, là aussi, s'en tirer très bien.

La question est donc : qu'est-ce qui n'est pas du ressort du singe ? Quelle fraction de notre temps exerçons-nous cette autre capacité ? Mon impression est que cette fraction du temps est extrêmement réduite. Je ne sais pas qui l'utilise. Les politiques quand ils décident d'une stratégie, est-ce qu'ils ont besoin de cette intelligence « non singe » ?

Mehdi Khamassi [30.35]

Peut-être que c'est même le cas dans n'importe quel corps de métier ; quand des spécialistes d'un sujet, d'un savoir-faire, extraient des règles un peu générales, et se rendent compte que dans plusieurs situations différentes, finalement, la solution qu'ils ont trouvée était du même type. Ils essaient alors d'abstraire l'essence de ce qui est commun à ces situations. Je ne sais pas si c'est quelque chose qui est mis en œuvre par des algos comme ChatGPT dans le domaine du langage à l'heure actuelle.

Michèle Sebag [31.05]

Attends, sur la capacité de généraliser, on peut leur faire confiance. Parce qu'encore une fois, dès que tu as un univers et assez de données, l'IA / ChatGPT sait généraliser de manière efficace et convaincante. Là où tu as besoin d'autre chose, c'est quand tu veux généraliser au-delà des données. Ou bien quand tu veux accéder aux autres, disons, barreaux de l'échelle cognitive, typiquement le raisonnement, des interventions, ou du contrefactuel. Pour ce qui concerne ces capacités, je ne pense pas qu'il [ChatGPT] en soit encore là. Mais je pense que ce sont clairement les nouvelles frontières. Or de manière très intéressante, parmi les systèmes qui sont capables de prédire, donc de répondre à ce qui est comme d'habitude, il y en a une fraction infime, de mesure nulle, qui est capable de répondre à ce qui se passerait en cas d'intervention. Et le même taux de chute énorme s'observe entre les systèmes qui sont capables de prédire ce qui se passerait en cas d'intervention, de préconiser des interventions, et les systèmes qui sont capables de raisonnement contrefactuel tels que « voici ce qui se serait passé si on n'avait pas fait ça ».

En résumé, à chaque barreau de cette échelle qui fait intervenir l'imagination, et le fait de jouer avec la réalité comme si on pouvait revenir sur le passé et refaire le film, je pense (1) que ces capacités sont énormes pour la conception, (2) qu'on s'en sert assez peu dans la vie de tous les jours, (3) que c'est la nouvelle frontière pour les systèmes d'IA, et en particulier (4) que c'est quelque chose d'essentiel si on veut que les systèmes se mettent (au lieu de répondre à nos questions qui font ce qu'elles peuvent), si on veut que le système puisse répondre aux questions qu'on adresserait à un oracle, c'est-à-dire « que faudrait-il faire pour obtenir tel résultat ». Donc, là, il y a encore énormément de marge. Mais avec les possibilités, les dangers augmenteront dans des proportions exponentielles.

2. Apprentissage de modèles internes

Mehdi Khamassi [0.08]

Alors justement, je suis ravi, ça m'amène à la question que je voulais poser d'un point de vue recherche, technique : pour faire un raisonnement contrefactuel – e.g., qu'est-ce qui se passerait si on

arrivait dans telle situation qui n'a pas été celle qu'on a rencontrée – il faut pouvoir se construire ce qu'on appelle un modèle interne du monde…

Michèle Sebag [0.26]

Oui.

Mehdi Khamassi [0.27]

… de façon à pouvoir faire de la simulation mentale…

Michèle Sebag [0.28]

Oui.

Mehdi Khamassi [0.29]

… donc pouvoir anticiper les conséquences des actions même quand on ne les a pas faites dans le réel. Et donc, pas seulement dans le domaine des algorithmes du langage, mais plus généralement des algorithmes d'IA en ce moment et d'apprentissage machine qu'on développe, j'ai l'impression qu'il y a cette tentation de beaucoup vouloir développer un gros réseau qui va construire un gros modèle interne. C'est le cas notamment dans un article de positionnement de Yann LeCun cette année, qui met l'accent sur l'apprentissage par un agent artificiel d'un grand et unique modèle interne qui lui permettrait de comprendre et de prédire ce qu'il entoure. Quelque part, ce que j'en comprends, c'est que pour lui toute l'intelligence, toute la clé, reposerait dans ce type de modèle interne à faire apprendre par l'agent. Mais d'un autre côté, j'ai envie de te rejoindre en faisant le lien avec un petit peu de cognition humaine : tu l'as dit, et je suis complètement d'accord avec toi, c'est montré en psychologie, la proportion du temps que les humains passent à utiliser leurs possibles modèles internes, à faire de la simulation mentale, à raisonner de façon contrefactuelle, à essayer d'anticiper ou de planifier avant d'agir, c'est un temps faible par rapport à beaucoup d'autres types de comportements qui sont beaucoup plus réactifs, beaucoup plus intuitifs, beaucoup plus rapides. Et en même temps il y a cette capacité qui a été démontrée de multiples fois : on peut poser des problèmes particuliers qui vont pousser les humains à exploiter cette capacité de simulation mentale, de raisonnement contrefactuel. Et on a l'impression, c'est en tout cas mon impression, qu'il y a un grand degré de distribution de l'information et des calculs dans le cerveau, et même de modularité quand même à un certain degré, qui font que je me dis intuitivement que faire un gros système, un gros réseau, je ne vois pas forcément très bien comment ça pourrait fonctionner. J'aurais plutôt envie de découper les choses [en plusieurs sous-réseaux correspondant à plusieurs modèles internes distincts]. Notamment, pour prendre un exemple concret, un gros réseau pourrait amener à des interférences, ce qu'on appelle en anglais du catastrophic forgetting. Alors est-ce que tu penses que c'est quand même une direction prometteuse ? Qu'est-ce que tu en penses ? Ou bien est-ce que tu penses qu'il faudrait faire autrement ?

Michèle Sebag [2.24]

Mais le problème c'est que… je pense que, à ce niveau, le mieux serait d'illustrer la vidéo par une image du système de LeCun. Parce que moi, ce que j'en ai retenu, c'est non pas que c'était un

gros réseau neuronal, un sac de spaghetti, qui fait tout, mais qu'il y avait une certaine modularité, en particulier au niveau des modules qu'il argumente en s'inspirant des aires du cerveau. Donc je pense qu'il est modulaire, qu'il n'est pas monolithique. Et je pense aussi peut-être que la créativité du cerveau vient exactement du fait que le cerveau comprend plusieurs joueurs. Donc je vais ici parler, je t'en avais parlé, de la vision de Jurgen Schmidhuber, qui est, je pense, historiquement, le premier découvreur des réseaux neuronaux récurrents, etc., mais qui n'a pas été reconnu par le prix Turing. Et donc lui propose cette vision de la créativité consistant à dire nous avons sous notre crâne deux agents : un qui essaie de former des lois, donc de compresser l'information, pour obtenir une vision claire et zen de ce qui s'applique ; un code en somme ; et l'autre agent sous le crâne essaye de déstabiliser le premier, essaie de produire des choses qui vont échapper à la codification du premier. Et de ce jeu dynamique entre les deux agents, un qui compresse et qui dit « tout est comme d'habitude », et l'autre qui dit « voici du nouveau », eh bien c'est cela que Schmidhuber recommande comme façon d'obtenir quelque chose de créatif. Donc je pense que quelque chose qui serait parfait, je pense, quelque chose qui manque dans la vision de Schmidhuber, que je trouve très très belle, c'est un troisième joueur qui est exactement l'effet social. C'est-à-dire : il existe beaucoup de disruptions, de façons de surprendre le joueur qui compresse, mais toutes ne sont pas également bonnes. Si on pouvait faire rentrer dans la boucle un être humain, ou un partenaire, qui permette d'évaluer ce qui est plus intéressant – ici, on est pour le coup dans du mou, qu'est-ce qui est plus intéressant que quelque chose d'autre ? C'est extrêmement subjectif ! –, mais je pense que là, avec une part du cerveau qui trouve le code de compression, nous avons la généralisation. Avec une part du cerveau qui sait surprendre la précédente nous pouvons sortir du cadre. Avec l'effet du troisième partenaire nous avons le fait que toutes les sorties du cadre ne sont pas également belles, ou élégantes, ou recommandables, etc.

Importance de la décomposition et de la cognition sociale

Mehdi Khamassi [6.38]

Alors je trouve que tu touches du doigt avec l'aspect social quelque chose de vraiment important. Et peut-être que ça me donne envie de préciser, sur l'aspect modulaire par rapport à Yann LeCun, et en tout cas ma compréhension de sa proposition : je suis complètement d'accord sur le fait qu'il y a une modularité dans les processus. Et puis il a ses modules : le configurateur, le simulateur, celui qui va calculer le coût. Et puis, quelque part, dans les generative adversarial networks (GANs ; en français, réseaux génératifs adversariaux), il y a l'idée qu'il y en a un qui va apprendre un modèle du monde, et puis l'autre qui va être le processus de génération de quelque chose de différent pour essayer d'attaquer l'autre, et qui va d'ailleurs aider, à la fin, ce grand modèle à être appris. Mais au fond il y a deux choses que je voudrais mentionner chez l'humain. Et encore une fois, ceci est dans ma compréhension et peut-être que je me trompe ; peut-être que ce n'est pas ça ce qu'il faudrait faire sur une machine, puisqu'il y a des capacités de mémoire et de calcul peut-être potentiellement quasi illimitées.

Mais chez l'humain en tout cas, on a l'impression que le module qui apprendrait un modèle interne pourrait être découpé, contextualisé, de sorte que dans une situation donnée, pour savoir ce qu'il faudrait faire, je ne vais pas avoir besoin de calculer tout ce qui serait possible de faire dans toutes les situations. Donc la situation donnée et le contexte vont pouvoir orienter la façon dont je vais raisonner. Par exemple, je ne vais pas raisonner de la même façon si je suis dans un contexte social ou

non social. C'est intéressant. Quand je cherche à comprendre qu'elle peut être la conséquence de mon action lorsque je fais un geste [je réalise qu'] il peut se passer des conséquences non sociales (e.g., je fais tomber un objet qui est sur la table), et il peut se passer des conséquences sociales (e.g., un autre être humain va réagir, va interpréter mon geste et s'adapter en conséquence). Et même si une partie des mécanismes dans le cerveau humain semblent communs, il y a en même temps des choses qui ont l'air très différentes et qui font que l'humain va surinterpréter ou suranticiper les réactions sociales possibles des autres partenaires.

Et donc, du coup, pour moi ça suggère qu'il y a vraiment peut-être un découpage, et que ce n'est pas un seul et même modèle qui va faire qu'on va avoir la même tendance à faire du raisonnement contrefactuel dans tous les contextes, en intégrant tous les types de connaissances qu'on a pu acquérir. Au contraire, il semble qu'on peut découper, et que peut-être ça simplifie dans plein de situations. C'est pour ça que je suis tenté de penser que cela peut être une solution intéressante de pousser encore plus dans cette direction. Mais peut-être qu'au fond c'est ce que Yann LeCun fait, et que je n'ai pas forcément saisi cet aspect. Ou peut-être que même si ce n'est pas ce qu'il fait, les chercheurs dans la communauté en IA, toi, Yann et d'autres, vous pensez peut-être que ça peut quand même bien fonctionner en ayant ce grand réseau.

Michèle Sebag [9.02]

Mais, d'un côté, tu as le fait qu'effectivement le monde est différent pour l'être humain et la machine. Et les bonnes solutions pour l'un ne sont pas nécessairement les bonnes solutions pour l'autre. D'un autre côté, pour le moment notre modèle indépassable, c'est le modèle humain. Mais, la nature étant économe, elle peut avoir intérêt à faire jouer plusieurs rôles à un même module. Du point de vue de la conception d'un système, on n'est pas sûr du tout que ça simplifie la conception du système. Donc peut-être une chose est de résoudre, et peut-être une autre chose est de construire celui qui résout.

Pour le moment, si on reprend ces vieilles images, on pourrait dire que le cerveau humain est plutôt un bazar alors que l'architecture proposée par Yann LeCun est une cathédrale. Mais je pense que l'inspiration est la même, c'est-à-dire la multiplicité de rôles, et le fait que l'essentiel est dans leur interaction. Le fait que nous n'évaluons pas tous les aspects d'une situation est aussi très lié à nos limitations en mémoire. C'est probablement aussi un leg darwinien [i.e., un héritage de l'Évolution], qui fait que si tu réfléchis trop longtemps ça peut être dangereux. Mais le concepteur du système n'a pas cette limite. Donc le fait de commencer par l'équivalent d'une cathédrale peut être une très bonne façon de faire.

Je vais me permettre de te poser à toi une question : toi qui as vu les différents modules de l'architecture de Yann, est-ce que tu vois des manques ?

Mehdi Khamassi [11.18]

Alors, il y a beaucoup de modules et de noms qui résonnent avec des modules qu'on utilise dans les architectures cognitives qu'on construit comme modèle de ce qu'on comprend, en partie, de la cognition et de l'apprentissage chez l'humain. L'idée d'un prédicteur, l'idée d'un modèle qui va sélectionner les actions, l'idée d'un module qui va calculer les coûts. Donc, il y a beaucoup de choses qui me semblent similaires. Je pense que cette distinction des contextes sociaux et non sociaux, le fait de distinguer le type d'agents avec lesquels on interagit – il y en a qui vont être animés, qui vont pouvoir faire des calculs à l'avance, et essayer de m'influencer dans l'autre sens, et ce n'est pas la même chose que d'interagir avec des objets non-animés –, je ne suis pas sûr d'avoir vu une telle dis-

tinction [dans l'architecture proposée par Yann LeCun]. Et ça me semble, dans ma compréhension de la cognition humaine, avec aussi mes limites, car je ne suis pas spécialiste de l'aspect social. Mais ça me semble quand même être un point clé des différences de raisonnements chez l'humain. Ensuite, il y aurait tellement de choses à réfléchir, à discuter. Il faudrait aussi que nous posions des questions à Yann lui-même. Mais ce point-là me semble important et je ne l'ai pas vu assez développé, peut-être.

Distinction conscient/inconscient

Michèle Sebag [12.24]

OK. Est-ce que par exemple il y a une notion d'inconscient ?

Mehdi Khamassi [12.32]

Je ne me rappelle plus. Il me semble qu'il cite notamment l'idée de Global Workspace (en français, l'Espace Global de Travail) de Stanislas Dehaene. C'est l'idée qu'il y a un certain nombre de processus qui ne sont pas au niveau conscient, mais qu'à un moment donné dans certaines situations, il peut y avoir une entrée en résonance entre beaucoup de processus différents – notamment lorsqu'il se passe des choses qui nécessitent qu'on réfléchisse sur la situation –, et qui peuvent faire qu'on passe [qu'on bascule] dans le domaine conscient. Maintenant, je ne crois pas que Yann ait la prétention d'aller aussi jusqu'à dire qu'il fait un modèle de la conscience, ni quoi que ce soit de tel, justement, heureusement.

Renouer avec les pères fondateurs de l'IA

Daniel Andler [13.03]

Je peux là jeter une observation, une question pour Michèle. Je n'ai pas lu en détail la proposition de Yann. Mais j'ai écouté plusieurs de ses exposés récents, et je vois, disons, l'orientation intellectuelle générale. Ce qui est quand même très frappant, c'est que cet homme qui a donc été une sorte de héros, non seulement intellectuel mais institutionnel – pour donc avoir fait valoir finalement l'orientation connexionniste aux dépens de l'orientation symbolique, qui ne voulait pas le savoir, qui se considérait comme très supérieure, etc. –, le voilà qui revient, me semble-t-il, un petit peu à ce qui était la perspective des pères de l'IA. Je voudrais demander à Michèle ce qu'elle en pense, car elle connaît bien l'histoire de l'IA. Or, cette perspective des pères de l'IA était une perspective essentiellement symbolique. Mais pour ne pas dire « ah oui, finalement, je reviens tout simplement au symbolique que j'ai combattu toute ma vie », il [Yann LeCun] dit « oui, je reviens en symbolique, mais en passant par la cognition humaine. Donc je m'inspire de la cognition humaine. » Or, c'était l'inspiration des pères. C'était l'inspiration de Simon, cette idée de faire à la fois de l'humain et de la machine, l'un étant l'envers de la médaille de l'autre, etc.

Donc je trouve ce mouvement de Yann comme reconnaissant au fond que le deep learning (en français, l'apprentissage profond) actuel, dont le connexionnisme mené vraiment à son extrême, a atteint en quelque sorte ses limites. Si on veut arriver à une véritable intelligence humanoïde, il faut

une nouvelle idée. Cette idée, au fond, c'est de revenir à l'ambition des pionniers, mais en passant maintenant sérieusement par la case sciences cognitives, parce qu'il y a 60 ans la case sciences cognitives était essentiellement vide, et maintenant elle est pleine de renseignements. Il faut exploiter ces renseignements pour réussir le projet. Est-ce que cette interprétation de l'histoire te paraît plausible, et sinon de quelle manière faudrait-il la corriger ?

Michèle Sebag [15.21]

Elle me paraît plausible, mais je vois assez bien la position des pères fondateurs, qui est elle-même différente de la position de Turing, qui est arrivée avant, et sur laquelle je reviendrai dans une seconde. La différence entre Yann LeCun et les pères fondateurs en termes de séquences et priorités est, je pense, que les pères fondateurs avaient le schéma « compréhensible, donc efficace », tandis que je pense que Yann LeCun a le schéma « efficace, et ensuite compréhensible si possible ». Mais le fait qu'il est plus facile d'essayer de comprendre un système efficace, même si ouvrir une boîte noire est une entreprise difficile, c'est plus facile que de faire marcher quelque chose de compréhensible. Et je pense que ceci n'est pas si loin, si on veut réfléchir aux pères fondateurs, n'est pas si loin de la démarche humaine. Il y a ces propos, que j'aime toujours beaucoup, de Brouwer, qui dit « j'ai la solution, il me reste à la démontrer ». Donc il y a bel et bien deux voies : une qui consiste à trouver, et une qui consiste à expliquer ou prouver. Et je pense que la nouveauté de Yann LeCun, je pense, à qui évidemment on ne peut pas reprocher sa timidité à braver les idées reconnues, consiste à dire : « OK, maintenant que ça marche, on va essayer d'aller plus loin, et en particulier de lui donner ses lettres de noblesse en termes de cognition humaine et/ou en termes d'explication, sinon du modèle, du moins de la manière dont l'IA fonctionne ».

Mais le fait de faire intervenir la cognition humaine était quelque chose qui était loin des pensées de Turing. Turing avait une limite dure, et il disait que « la limite ne va pas être les données ». Il avait montré ça avec des estimations rapides et grossières. La limite va être de créer des programmes intelligents : « moi, en tant que bon programmeur, je peux écrire tant de lignes de code par jour. Et comme je vois dès aujourd'hui que ça ne va pas suffire, la seule solution, c'est que la machine apprenne elle-même ». Et il évoque quelque chose pour faire apprendre à la machine, quelque chose qui est peut-être très bas ou peut-être très tôt dans la cognition, qui est typiquement Pavlov. À force d'émettre des signaux de type « c'est bien », « ce n'est pas bien », on va réussir à faire comprendre à la machine ce que l'on veut, et à l'obtenir.

Daniel Andler [19.12]

Je te suis. Merci. Oui, oui, tout à fait.

Mehdi Khamassi [19.17]

OK, alors, de mon côté, il me reste deux petites questions qui sont vraiment d'un autre domaine. Mais si vous voulez approfondir cette discussion, je laisserai la place à une autre question de Daniel. Juste pour les annoncer : une question est sur des applications de l'IA qui puissent être utiles à la société. Je pense que c'est important d'y réfléchir, d'avoir des exemples, et je sais que Michèle, de ton côté, tu fais des choses en ce sens. Donc ça serait vraiment intéressant d'en parler. Et l'autre question porte sur la soutenabilité en termes de technologie. Est-ce que vous, ou toi Daniel, vouliez d'abord que l'on creuse une autre question ? Ou on y va ?

Daniel Andler [19.51]

Je pense que tu as raison. J'avais juste, comme je suis toujours très curieux d'écouter Michèle, et mesurant une fois de plus l'extension de son savoir et de sa sagesse, j'ai envie de lui poser cette question banale, ici finalement, mais peut-être pas tellement importante. La question est la suivante : est-ce que l'intelligence artificielle a une unité et est-ce que l'intelligence artificielle se distingue clairement de l'informatique ? Il s'agit d'une question de positionnement, finalement, disciplinaire de l'intelligence artificielle. Il y a des gens qui disent que l'expression devrait être bannie tout simplement, parce que ça veut dire 100 choses différentes pour 100 personnes différentes, ça ne sert à rien, ça ne fait que tout embrouiller. Ou bien est-ce que « non, quand même » ? Moi j'ai tendance personnellement à penser qu'il y a une certaine unité de l'IA, finalement, et qu'on a besoin de ce terme. Mais du coup, j'ai évidemment du mal à tracer une frontière entre l'IA et l'informatique, ou l'informatique simple, ou quelque chose comme ça. C'est un peu une question méta, qui n'a peut-être pas tellement sa place dans cette interview. Enfin… ça dépend… c'est à Michèle de savoir si elle veut développer ou pas.

Michèle Sebag [21.05]

Ce n'est pas à toi que je vais apprendre que la science est sociale et que les noms des choses emportent des positions de pouvoir. Donc le fait de dire ou non que l'IA n'est autre que l'informatique revient à ruiner un pouvoir. Tu as pour le moment beaucoup de gens qui, en voyant le gâteau de l'IA, découvrent qu'ils faisaient de l'IA depuis qu'ils étaient tout petits, et qui rejoignent le wagon en apportant des compétences extrêmement précieuses, par exemple, en termes de preuves ou de certifications. Ceci revient à réutiliser l'informatique du siècle dernier, disons, pour polir, pour habiller, pour rendre plus digeste l'IA. Est-ce qu'il y a une signature des gens qui font de l'IA ? C'est là où je ne sais effectivement pas s'il y a une différence ou non avec l'informatique. Une signature des gens qui font de l'IA consiste typiquement à croire que tout est possible. Ou plutôt, c'était le cas. (rires) Je pense que le reste de l'informatique est beaucoup plus conscient des limites imposées par l'infrastructure. Ceci va dans le sens de la question que je ne veux pas chercher à éviter, la question de Mehdi sur la soutenabilité.

3. Applications de l'IA et soutenabilité

Mehdi Khamassi [0.08]

De mon côté, il me reste deux petites questions qui sont vraiment d'un autre domaine. […] Juste pour les annoncer : une question est sur des applications de l'IA qui puissent être utiles à la société. Je pense que c'est important d'y réfléchir, d'avoir des exemples, et je sais que Michèle, de ton côté, tu fais des choses en ce sens. Donc ça serait vraiment intéressant d'en parler. Et l'autre question porte sur la soutenabilité en termes de technologie. […]

Michèle Sebag [21.05]

[…] Une signature des gens qui font de l'IA consiste typiquement à croire que tout est possible. Ou plutôt, c'était le cas. (rires) Je pense que le reste de l'informatique est beaucoup plus conscient

des limites imposées par l'infrastructure. Ceci irait dans le sens de la question que je ne veux pas chercher à éviter, la question de Mehdi sur la soutenabilité.

Telle qu'elle est, l'IA n'est pas soutenable. Est-il pensable d'avoir un point d'arrêt ? Je ne crois pas. Si tu veux, tant que les acteurs ont le choix de leur décision, alors j'ai l'impression que c'est comme sur l'île de Pâques : on va faire des IA de plus en plus grosses, jusqu'au moment où toutes les ressources auront été mangées, ou en tout cas une partie des ressources utiles à la survie. Et je ne vois pas qu'on puisse éviter cette bataille, qui est à la fois pour les ego, à la fois pour le commerce, à la fois pour la science. Si tu sais que tu peux, la règle [que les gens ont l'air de suivre] est à peu près « tu dois ». Donc je ne vois pas quelle force présentement pourrait imposer un plafond à la somme d'énergie qui est développée et qui est employée pour développer une IA. Je pense que ce n'est pas forcément faisable dans un contexte où les lois du capitalisme autorégulé s'appliquent.

Mais bon tu vois ça c'est à propos de l'IA, mais les mêmes problèmes s'appliquent, sauf que c'est beaucoup moins notre tasse de thé, pour les tanks (bruit), et pour tout ce que tu veux : plus gros, plus beau, mieux. Et on a toujours l'impression de faire un prototype, ce qui permet de croire qu'après celui-là on va s'arrêter. Sauf que personne ne s'arrête.

La formation à l'éthique est-elle la solution ?

Mehdi Khamassi [3.22]

Au fond, tu l'as souligné, beaucoup de gens, ou un certain nombre de gens, croient que tout ce qu'il est possible de faire, on doit le faire. Et quelque part, là on est vraiment dans un questionnement éthique personnel. D'où la question de la formation à l'éthique des ingénieurs, des scientifiques ; formation qui permettrait d'ouvrir au fait de ne pas forcément croire que c'est ce qu'il faut penser [que tout ce qu'il est possible de faire, on doit le faire]. Il y a justement des fois où on pourrait [faire], mais où au fond ce ne serait pas souhaitable pour la société. Et on s'en rendrait compte si on prenait le temps de réfléchir et d'aller dans cette direction, celle de l'éthique. Après, ça ne veut pas juste dire qu'il faut fermer et puis s'arrêter et passer à autre chose. Mais il s'agit éventuellement de trouver une autre voix ou de dire « tiens, l'IA, si on la poussait vers une question d'utilité sociétale, vers une direction plus soutenable, on pourrait continuer d'aller toujours vers du possible, et de toujours pousser le plafond ». Donc ça peut être fascinant, ça peut être stimulant. Mais ça devrait peut-être se faire en évitant des directions où on se rend bien compte qu'il y a de fortes chances que ça ne soit pas soutenable.

Cette formation à l'éthique [notamment des ingénieurs et des scientifiques, mais pas seulement], est-ce que c'est le Graal ? Est-ce que c'est ça ce qu'il faut faire ? Ou finalement cela revient à penser uniquement à l'échelle individuelle ? Et en plus cela prendrait des décennies pour intégrer ça pour les prochaines générations ? Est-ce que c'est illusoire de penser qu'on peut progresser là-dessus ou que c'est ça la solution ?

Michèle Sebag [4.43]

Comme d'habitude tu poses des questions auxquelles je n'ai pas la réponse. Mais tu vois : est-ce que la formation à l'éthique va résoudre le problème ? Je ne le pense pas, pour une raison qui est que : le fait que de mauvaises choses ne se produisent pas ne dépend pas du fait qu'une majorité de

gens soient éthiques, mais du fait que personne ne soit fortement non-éthique, si j'ose dire. Et ça, je pense que c'est irréaliste. Je pense que l'Histoire nous montre que des contextes dans des civilisations hautement évoluées ont pu produire des choses qui soient détestables pour les autres ou détestables pour eux-mêmes, pour elles-mêmes.

Donc l'éthique, c'est beaucoup mieux que rien. C'est nécessaire. Est-ce que ça va être suffisant ? Je ne crois pas. Est-ce qu'il faudrait qu'il y ait des centres comme ceux où on joue avec les virus de la peste et du choléra pour élaborer des modèles d'IA hautement sulfureux ? Ça pourrait être très très intéressant.

Utiliser l'IA pour aider à résoudre des problèmes de société (exemples : soutenabilité, santé, emploi, organisation du CNRS)

Michèle Sebag [6.13]

Est-ce qu'on pourrait se servir de l'IA pour produire une société plus soutenable ? Là encore le problème est dans les effets. Tu vois, par exemple, il y a des études, notamment une des études sur lesquelles on est en train de travailler : est-ce qu'il y a un impact causal entre les problèmes de santé et la distance aux champs et aux pesticides ? Donc voilà [une question à laquelle on peut apporter] une réponse factuelle. D'ailleurs, obtenir les données relève d'un parcours du combattant. Un exemple d'un autre type d'études, qui est beaucoup plus proche des interventions, est le suivant : quand on est en train d'essayer d'agir, par exemple, pour les établissements publics pour l'emploi, Pôle emploi, par exemple, alors la subtilité des questions éthiques qui se posent est abasourdissante. Il y a typiquement des biais dans les données. Sans même avoir besoin d'un phare, tu vas voir par exemple que les rouges sont plus payés que les bleus, ou que les rouges vont travailler plus loin que les bleus, etc. Et maintenant, là-dessus, qu'est-ce que tu dois faire ? La plateforme peut se retrouver à parfaitement amplifier des biais qui existent.

Quelqu'un racontait que du point de vue des chauffeurs UBER – où pourtant tu peux te dire « là, franchement, les préjugés sur les chauffeurs devraient ou pourraient être assez peu présents » –, mais le fait est qu'il y a une rémunération plus élevée pour les chauffeurs hommes, du fait qu'ils vont aller dans des quartiers chauds la nuit, et que ces courses-là paient bien. Donc voici une possibilité qui est ouverte par la plateforme, qui s'insère dans un monde qui est ce qu'il est, et qui a un effet directement mesurable. Qu'est-ce qu'on peut faire avec nos systèmes de recommandation sur Pôle emploi ?

Je vais prendre maintenant l'exemple d'un projet sur lequel je n'ai pas travaillé, mais sur lequel j'ai envie de travailler depuis un certain temps, mais qui est peut-être trop conflictuel. Suppose par exemple qu'on propose à l'IA d'élaborer une politique de l'impôt. Qu'est-ce que tu lui donnes en entrée ? Tu lui donnes les secteurs que tu veux développer. Tu lui donnes les données. Tu te souviens ? Piketty, Landais et Saez avaient fait un livre dans lequel il y avait un simulateur merveilleux, du style « testez vos idées idiotes sur ce qu'on devrait faire avec l'impôt ». Ce n'est pas exactement ça, mais bon. Et donc, tu voyais que les trois ou quatre premières idées qui te venaient à l'esprit avaient un effet peanuts, en d'autres termes un effet totalement misérable. Ce qui manquait dans le système de Piketty, et ce qui serait absolument super, mais peut-être aussi terrifiant, serait de savoir comment les gens auraient réagi à ces initiatives. Tu vois, tu as un système. Si ma mémoire est bonne, ils avaient

découpé la population française en 80 000 types. Donc c'est un boulot de bénédictins. (rires). Et on voyait très très finement quel serait l'impact à court terme des décisions qu'on peut prendre en bougeant les curseurs par-ci par-là. Mais évidemment la société ne serait pas restée immobile. Donc tu devrais bel et bien concevoir le système de l'impôt comme un jeu de compétition. Et c'est là où on a absolument besoin de dire « si je fais ça, les gens vont faire ci », de façon à ce que le résultat de tes bonnes idées ne soit pas pire que l'état initial.

Qu'est-ce qu'on pourrait faire là ? Mon impression est que pour arriver à commencer à toucher du bout du doigt le problème, tu serais forcé de déconstruire tant de relations de pouvoir, que je ne suis pas sûre d'en voir un jour le bout.

Je vais prendre un autre exemple, qui à mon avis intéresse assez peu de gens, qui est la manière dont la section informatique du CNRS a été séparée entre sections 6 et 7. Bien, bien, bien, bien, bien. Pardon pour ceux qui ne s'intéressent pas à ce problème, qui est à la vérité assez confidentiel. Pour résumé, on a un machin qui s'appelle l'informatique, qui devient une section très grosse, et qu'il s'agit de splitter (en français, de diviser) pour une quantité de raisons qui sont bien fondées. Le split, i.e., le fait de découper les deux sections, est un enjeu de pouvoir énorme. Tu avais des méthodes techniques : si tu regardais l'amplitude de l'ancienne section informatique, tu pouvais appliquer les propres outils de l'informatique, typiquement de décomposition de graphes, pour arriver à voir où faire la coupure et qui fait le moins de rupture. Si naïf que ça puisse paraître, je l'ai proposé. On m'a dit : « ah, c'est intéressant ». Mais, tu vois, jamais ! Même en rêve !

Donc, je pense qu'il faut concevoir directement les outils d'IA dans un contexte de pouvoir. Le fait de ne pas les placer là donne à mon avis lieu à des discussions assez théoriques.

Mehdi Khamassi [13.20]

Ce que je retiens particulièrement de ce que tu dis, c'est à la fois que tu as touché à des enjeux que l'on pourrait presque considérer comme relevant de la psychohistoire, comme Asimov dans Fondation, c'est-à-dire modéliser [des problèmes de société] comme un système dynamique, ce qui revient à prendre en compte la réaction des gens à tel ou tel changement. (bruit) Je suis désolé pour le bruit des travaux qui ont démarré. Je ne sais pas si vous les entendez ou si ça va. Et donc là j'y vois quelque chose qui m'a toujours stimulé, d'une part quand j'étais en école d'ingénieurs et que j'avais mes premiers cours sur l'IA, et d'autre part [quand j'étudiais] la modélisation statistique : je rêvais de contribuer à mettre au point des algos d'IA qui puissent nous aider à faire sens du monde dans lequel on vit, contribuer à mieux comprendre la société, pour essayer de l'arranger de façon à ce qu'on puisse y vivre mieux ensemble, quelque part. Et j'ai l'impression qu'il y aurait la possibilité, suite à cette discussion, de mettre encore plus le projecteur sur ces voies possibles, qui peuvent être stimulantes intellectuellement, en termes de développement et de conception de systèmes d'IA, et pourraient en même temps contribuer à des applications sociétales qu'on espère positives, qu'on espère utiles.

D'un autre côté, ce que je retiens de ce que tu viens de dire à la fin, c'est aussi que se contenter de faire ce travail-là dans son coin, de façon presque monodisciplinaire, mais sans compréhension des enjeux de pouvoir, des problématiques sociologiques, et donc de plein de choses qui ont été comprises et mises en lumière dans d'autres disciplines, on pourrait presque se retrouver dans une impasse. Donc il y a absolument besoin de cet échange.

Michèle Sebag [14.45]

Oui.

Souligner les potentiels positifs de l'IA sans négliger les possibles effets négatifs

Mehdi Khamassi [14.46]

En tout cas, je trouve que mettre l'accent sur les applications possibles de l'IA qui soient utiles pour la société, c'est quelque chose qui me semble important. On entend beaucoup en ce moment qu'un certain nombre de jeunes ingénieurs sont en quête de sens. Peut-être que ça pourrait être stimulant de leur côté.

Michèle Sebag [15.06]

Oui.

Mehdi Khamassi [15.07]

Après, est-ce que c'est une façon de sans arrêt mettre le projecteur sur ce qui peut être positif, alors que finalement cela masque un petit peu la forêt et tout plein d'autres effets de bord qui peuvent être négatifs ? Est-ce que c'est un risque ?

Michèle Sebag [15.23]

Si tu veux, d'un autre côté, la forêt, on y est. Les foultitudes d'effets négatifs, on les voit.

Y a-t-il encore une tendance à aller vers des IA de plus en plus autonomes ?

Daniel Andler [15.32]

Alors je ne sais pas ce qu'en pense Michèle, j'enchaîne sur une autre question qui me venait en écoutant Michèle d'une manière qui me paraissait vraiment tout à fait pertinente : y a-t-il encore une tendance en IA – moi, il me semblait que oui, mais peut-être que je me trompe – à essayer d'aller vers des IA de plus en plus autonomes ? Donc d'avoir vraiment des intelligences artificielles qui seront meilleures que nous parce qu'elles voient plus de données, elles traitent plus rapidement, elles ne sont pas fatiguées, etc., etc., tout ce qu'on dit à ce propos ? Et donc le but serait de faire des IA de plus en plus autonomes. Mon sentiment est qu'il ne faut surtout pas faire ça, justement, parce que l'IA intervient dans un monde humain, avec des normes humaines, des effets pervers donc complexes de rétroaction sociale, etc. Donc il ne faut surtout pas aller vers des IA de plus en plus autonomes. Qu'est-ce que tu en penses, Michèle ?

Michèle Sebag [16.32]

Une IA autonome en contact avec des humains, c'est le point dont on parlait au début. Et le fait d'avoir des entretiens avec, disons, des métahumains ou bien des instructeurs qui permettent de

décharger la pression et de redevenir centré me paraît souhaitable. Pour des IA qui ne sont pas en interaction avec des humains, est-ce que l'autonomie est bien ? Ici me reviennent à la mémoire les propos de Stuart Russell, qui disait quelque chose du style : « Si vous donnez à une IA la fonction à optimiser, elle est parfaitement autonome, mais c'est le début de la fin des haricots, parce que si l'IA sait quelle est la fonction à optimiser, elle va pouvoir l'optimiser mieux que vous. Or malheureusement les chances sont importantes pour que ce ne soit pas la bonne fonction à optimiser [i.e. que l'humain se soit trompé sur le choix de la fonction]. Donc là on obtient une dystopie grandeur nature. Donc je suis assez d'accord avec toi : nous avons trop peu compris [quelles sont les fonctions à optimiser pour le bien des sociétés humaines], et nous sommes trop conscients des risques de dérapage, pour qu'on s'attaque à l'autonomie. Ça peut être un objectif. Mais dans la pratique il me semble que c'est vraiment trop dangereux, ou inapproprié.

Mehdi Khamassi [18.23]

C'est très intéressant. Ou en tout cas une autonomie en laboratoire, bien contenue, pourrait peut-être être utile justement pour faire réfléchir sur jusqu'où va la fonction de coût qu'on a défini au préalable, et faire réfléchir sur les fonctions de coût, dans la mesure où elles nous aideraient à réfléchir sur nous-mêmes, sur la façon dont on définit les critères.

Règlementations européennes sur l'IA

Michèle Sebag [18.40]

Peut-être qu'il y a encore un point à discuter : tu m'avais dit « essaie de voir pour les actes européens. » Donc j'ai fait mon travail hier soir.

Mehdi Khamassi [18.54]

Merci beaucoup ! (rires)

Michèle Sebag [18.56]

Je ne l'ai pas étudié suffisamment, mais je suis tombée sur une perle que j'ai bien envie de partager avec vous. Le début du Data Act de 2022 dit en substance : « les données ne sont pas des biens rivaux. De la même manière que la lumière des rues ou une belle vue [peuvent être accessibles à toutes et tous], beaucoup de gens peuvent accéder à ces données en même temps. Et elles peuvent être consumées encore et encore sans impact sur leur qualité ou sans courir le risque que la source se tarisse. » Ce que je trouvais très très beau était le fait que les exemples qui sont donnés de biens non rivaux, comme les données, sont : la lumière des rues, or ça vient de l'État, et a scenic view (un paysage, en français), ça vient de la nature. Donc d'une certaine manière, les meilleurs exemples qui sont donnés de biens non rivaux sont (1) ce qui est organisé par l'État ou (2) ce qui organisé par la nature, dont le coût est assez peu perçu.

Et donc maintenant la question est : qu'est-ce qui se passerait si l'État lui-même faisait une IA ? Franchement, je n'en sais rien, mais j'aimerais beaucoup participer à ça. Le fait que l'IA, avec les en-

jeux de pouvoir des services de l'État, pourrait intervenir là, j'ai l'impression qu'il y a déjà eu quelques initiatives qui n'ont pas forcément été concluantes. Mais voilà, je voulais juste… Tu vois, cette phrase est tellement… Tu te dis que les gens qui ont écrit ça sont peut-être un peu à l'ouest.(rires)

Mehdi Khamassi [21.43]

Je ne sais pas si tu voulais réagir, Daniel. Moi je n'avais pas spécialement décelé cette phrase. Mais elle me fait penser, et c'est un peu dans le sens de l'interprétation que tu donnes, biens non rivaux, biens communs. Comme exemple de bien commun, il y a toujours pour moi la connaissance : quelque chose qu'on peut partager, sans déposséder les personnes qui font un don de connaissance. Donc quelque part c'est l'idée d'un bien commun universel, qui serait organisé par la collectivité, donc le public, et l'État y contribuerait. Au fond, il est important de savoir garder la mesure de cette importance et du fait que c'est un bien commun. Alors on peut toujours penser aux débats sur le fait de privatiser les profits et de collectiviser les pertes. Donc si finalement on ne mesure pas ce coût, et que c'est ça ce qu'on garde dans le collectif sous la chapelle de l'État, et que tous les autres profits par ailleurs sont privatisés, dans quelle mesure est-ce soutenable d'un point de vue économique ?

Ça me paraît aussi être vraiment un enjeu important. Mais voilà je n'ai pas réfléchi au-delà sur cette phrase, donc je ne suis pas capable d'en dire plus.

Daniel Andler [22.40]

Il va y avoir justement dans quelques semaines un colloque à l'Unesco consacré à ces questions-là, et à la question de préserver l'information comme common good justement (bien commun, en français), quelle régulation imposer, etc. TESaCo participe un petit peu à l'organisation, de manière un peu in extremis, de ce colloque. En tout cas effectivement ce sont des questions absolument centrales.

Je pensais à une chose : quand je veux une trottinette parce que mon vélo a crevé, il y a un système qui me permet de dire à 10 mètres près où est la trottinette la plus proche, et à quel niveau est sa batterie. Je me disais « quand même, c'est un luxe invraisemblable. Enfin il y a des choses plus urgentes, merde ! » On a un système GPS extraordinaire, qui me permet de repérer, moi petit bourgeois qui cherche à repérer ma copine au restaurant, et voilà : « La trottinette est à 15 mètres d'ici. Batterie à 50% ». Et alors je me suis fait la réflexion que le GPS n'est pas là pour ça, il est là pour des choses beaucoup plus importantes. Mais une fois que le GPS est là, évidemment, il peut aussi servir à cette chose absolument triviale qui consiste à repérer la trottinette la plus proche. Mais c'est moins simple que ça. Et je ne sais pas très bien parce que ce n'est pas du tout ma spécialité, je ne sais pas comment réfléchir à cette question-là. Le fait que le GPS sert aussi à toutes sortes de choses dont on pourrait vraiment se passer, et compte tenu du fait que le GPS coûte très très cher, est-ce qu'il n'y a pas quand même un effet pervers, là ? En l'occurrence, le fait que théoriquement le GPS soit là pour des choses absolument essentielles mais qu'on s'en serve pour des choses totalement inessentielles, que cet usage inessentiel finisse finalement par l'emporter sur les usages essentiels et par orienter en quelque sorte la dépense publique, ou la dépense des ressources de l'humanité, pour des choses essentiellement triviales.

Michèle Sebag [24.56]

Je suis d'accord avec toi. Mais je ne vois pas comment conduire l'argumentation. Par exemple, et je change de domaine, quand tu regardes les arguments qui sont évoqués pour la 5G, tu te rends compte que c'est du type « c'est très bien pour la chirurgie ».

Daniel Andler [25.15]

Oui, oui, tout à fait. J'ai participé à un groupe de travail sur la 5G. Tu as tout à fait raison. C'est un excellent exemple.

Michèle Sebag [25.23]

Donc tu as quelque chose qui est de l'ordre du storytelling qui fait que tout usage est justifié s'il en existe une version bien, c'est-à-dire un objectif louable. Et inversement le même storytelling peut servir à décourager toute initiative s'il en existe des conséquences terribles, enfin non souhaitables, indésirables. Donc mon impression est que le storytelling est au cœur des systèmes qui nous donnent à voir la réalité, qu'on réagit de manière assez, éventuellement, réflexe, que le nombre d'usages possibles rentables fondés sur la technologie est sans frein. Mais maintenant j'ai une bonne nouvelle dans ce monde : je pense que l'Académie des Technologies a émis un rapport dans lequel elle disait à peu près : « toute innovation ne doit pas nécessairement être financée ». Là tu vois, tu touches aux fondamentaux.

Daniel Andler [27.03]

Bien sûr. L'Académie des Technologies a justement pour mission de dégager une sagesse, justement collective [comme dans le nom TESaCo] sur les technologies. Donc là elle fait son boulot, elle est dans son rôle.

Michèle Sebag [27.18]

Cette phrase a demandé des négociations à la virgule. (rires) Mais tu vois, c'est aussi un pas.

Daniel Andler [27.28]

Oui.

Mehdi Khamassi [27.30]

Et en plus je trouve que dans ce que tu dis, Michèle, il y a quelque chose de très important et très beau, et finalement auquel on est en train de contribuer aujourd'hui, qui est dans notre position, d'essayer de donner à voir des applications possibles, des conséquences possibles, certaines peut-être indésirables, on doit alors y réfléchir, certaines positives, et en même temps en essayant d'enlever le maximum de storytelling, et simplement de donner à réfléchir. C'est peut-être une des meilleures choses qu'on puisse continuer à faire.

En tout cas vraiment merci beaucoup, Michèle, encore une fois c'est vraiment super, je suis ravi de cet échange.

Daniel Andler [28.06]

Oui vraiment merci.

Michèle Sebag [28.07]

Merci à vous.

Mehdi Khamassi [28.08]

Et puis on te souhaite plein de bonnes choses et on te dit à bientôt !

Michèle Sebag [28.12]

À bientôt j'espère !

L'apprentissage profond hier et demain

Audition de Yann LeCun

YANN LECUN

Yann LeCun est Professeur à New York University, où il a créé le Centre pour les Sciences des Données, et Scientifique en chef sur l'IA à Meta (anciennement Facebook). Initialement diplômé d'une école d'ingénieurs en France, l'ESIEE, et il a obtenu un DEA puis un Doctorat en Informatique en 1987 de l'Université Pierre et Marie Curie à Paris. En 2018, il a obtenu le Prix Turing avec Yoshua Bengio et Geoff Hinton, pour son travail pionnier en apprentissage automatique et sur les réseaux de neurones récurrents et profonds. Il a notamment inventé les réseaux convolutifs pour la reconnaissance d'images.

L'audition a été menée par Mehdi Khamassi et Daniel Andler

Première partie de l'audition : Aux origines de l'apprentissage profond

Mehdi Khamassi [0.08]

Bonjour Yann LeCun. Merci beaucoup de nous accorder du temps et d'accepter de répondre à nos questions pour le projet TESaCo, Technologies Émergentes et Sagesse Collective, de l'Académie des Sciences Morales et Politiques. C'est un projet dirigé par le philosophe Daniel Andler, spécialiste des sciences cognitives et s'intéressant depuis longtemps à l'intelligence artificielle. Malheureusement, il n'a pas pu être présent aujourd'hui, mais il n'a pas manqué de m'envoyer un certain nombre de questions. Ainsi, il contribue indirectement à cette discussion.

Tu es un spécialiste de l'intelligence artificielle (IA), initialement diplômé de l'École Supérieure d'Ingénieurs en Électronique et Électrotechnique (ESIEE) en France. Ensuite, tu as effectué un DEA et un doctorat en informatique que tu as obtenu en 1987 à l'Université Pierre et Marie Curie, aujourd'hui appelée Sorbonne Université, à Paris. Tu es reconnu comme l'un des pionniers de l'apprentissage automatique, des réseaux de neurones artificiels profonds.

Tu es aujourd'hui Professeur à l'Université de New York, où il me semble que tu as créé le Centre pour les Sciences des Données.

Yann LeCun [1.09]

Oui.

Mehdi Khamassi [1.10]

En même temps, tu occupes le poste de scientifique en chef pour l'IA chez Meta (anciennement Facebook). Et puis, chose très importante qui est une reconnaissance de l'importance de tes travaux, tu as reçu le prix Turing en 2018 aux côtés de Yoshua Bengio et Geoffrey Hinton, pour vos contributions théoriques, empiriques et autres dans le domaine des réseaux de neurones profonds.

Une de tes contributions majeures est le développement des réseaux convolutifs pour la reconnaissance d'images. Cela m'amène à te poser une première question d'ordre historique, car je trouve cela intéressant. Il me semble que du côté des neurosciences et de la perception visuelle chez l'humain, il y a aussi des processus similaires à la convolution, de l'intégration d'informations sur de petites surfaces de la rétine, suivie d'une mise en relation. À l'époque où tu travaillais sur ces sujets, existait-il déjà des liens explicites avec les neurosciences ?

Yann LeCun [2.07]

Oui, absolument. L'architecture des réseaux convolutifs est directement inspirée des travaux très classiques en neurosciences de Hubel et Wiesel au début des années 60, voire fin des années 50. Ces travaux leur ont d'ailleurs valu le prix Nobel au début des années 70. Par la suite, d'autres recherches ont essayé d'utiliser ce genre de concepts pour des modèles sur ordinateur, tels que les cognitrons et neocognitrons de Kunihiko Fukushima dans les années 70 et 80. C'était une époque où très peu de gens, particulièrement en Occident, travaillaient sur les réseaux de neurones. Donc, oui, il existe une filiation directe et une inspiration directe provenant des neurosciences.

Mehdi Khamassi [2.52]

C'est vraiment intéressant, ce lien et ces échanges qui ont pu se développer au fil du temps entre l'IA et les neurosciences, voire même les sciences cognitives peut-être en général, car il y a des questions liées à la psychologie cognitive et au développement de la cognition chez l'enfant, qui t'intéressent et sur lesquelles nous reviendrons un peu plus tard. Je pense que je te poserai d'autres questions, car c'est fascinant. On se demande souvent dans quel sens les inspirations circulent entre l'IA et les neurosciences. Et il me semble qu'elles peuvent s'entre-aider en fait assez fréquemment.

Peut-être une autre question historique : je me souviens d'une anecdote que tu as racontée. À une certaine époque, lorsque tu travaillais sur les réseaux de neurones et l'apprentissage automatique, qui était une époque où le domaine de l'intelligence artificielle était dominé par d'autres approches, notamment celles de l'IA symbolique. Et au fond, ce que tu faisais n'était peut-être pas particulièrement valorisé ou pris au sérieux. Je trouve ce genre d'anecdote très intéressant, car cela nous incite à de la modestie aujourd'hui et de ne pas dénigrer d'autres domaines qui peuvent ne pas être en vogue à un moment donné, mais qui pourraient se révéler prometteurs dans 10, 20 ou 30 ans. As-tu l'impression que ce type d'anecdote contribue à faire réfléchir les gens autour, ou bien est-ce que c'est un éternel recommencement ?

Yann LeCun [3.58]

Absolument. Je pense qu'il y a deux choses auxquelles il faut prêter attention. Tout d'abord, il est effectivement important de prêter attention à des idées qui sont peut-être un peu oubliées malgré leur valeur. Et puis surtout, il est crucial de ne pas se rallier aux modes qui tendent à créer une espèce

d'uniformisation. Aujourd'hui la mode est à «l'apprentissage profond» (*deep learning en anglais*). C'est normal, car ça marche. Et puis là, depuis 1 an ou 2, la mode est aux «transformers», aux LLMs (*Large Language Models* ; en français : les grands modèles de langage), etc. Beaucoup de gens placent de grands espoirs en ces approches. Pour ma part, j'ai été assez rapide à dire que ce sont des systèmes qui sont très limités, et ne seront pas suffisants pour atteindre le type d'intelligence qu'on observe chez les animaux et les humains.

En fait, il y a une succession d'histoires et d'engouements de la communauté de l'intelligence artificielle dès qu'un nouvel ensemble de techniques est découvert. Un certain nombre de personnes se disent alors : « ah, ça y est ! On a découvert le secret. On va arriver au niveau de l'intelligence de l'humain. » Il y a eu ce genre d'histoire périodiquement tous les dix ans depuis les années 50. Cela a commencé avec le «general problem solver» de Newell et Simon, puis avec le perceptron dans les années 60, d'une certaine manière.

Mehdi Khamassi [[5.27]

Et puis les systèmes experts, peut-être aussi à un moment donné.

Yann LeCun [5.29]

Les systèmes experts dans les années 80-90, et puis les réseaux de neurones à la fin des années 80. À chaque fois, les ambitions se renormalisent, mais les techniques inventées deviennent finalement des outils essentiels dans la boîte à outils des ingénieurs. C'est intéressant. Par exemple, tout le courant sur les perceptrons, ADALINE et autres du début des années 60, et fin des années 50, est un peu mort à la fin des années 60 à cause des limitations des classifieurs linéaires et autres systèmes à une couche. Et les gens ont simplement changé le nom de ce qu'ils faisaient. Au lieu de dire qu'ils essayaient de construire des machines intelligentes, ils se sont dit : « on va appeler ça maintenant avec des noms plus sérieux, comme «reconnaissance des formes statistiques» [c'est devenu les classifieurs linéaires], ou «filtrage adaptatif» ». En fait, toutes les techniques de communication, de modems, etc., étaient basées entièrement sur ces algorithmes qui avaient été inventés dans le début des années 60 dans le contexte des réseaux de neurones.

Mon propre parcours dans cette histoire commence vers 1980, alors que j'étais encore étudiant et que les bases du Deep Learning et des réseaux neuronaux étaient en train d'être posées. C'est un phénomène assez intéressant.

Mon histoire là-dedans démarre environ en 1980. J'étais encore étudiant et j'étais vraiment intéressé par le mystère de l'intelligence et puis par le fait que possiblement on pourrait faire en sorte que des machines s'auto-organisent et s'entraînent. J'ai découvert qu'il y avait eu des travaux là-dessus dans les années 50 et 60, mais qui n'existaient plus au moment où j'ai commencé à m'intéresser à ça. C'étaient des travaux qui avaient disparu. Et puis je me demandais bien comment j'allais pouvoir continuer à travailler sur la question : peut-être en faisant un doctorat ou un DEA [équivalent du Master aujourd'hui]. Et c'est là que j'ai rencontré Daniel Andler, entre autres, qui à l'époque était au CREA [le Centre de Recherche en Épistémologie Appliquée], une espèce de petit labo pirate…

(rires)

… dans les locaux du ministère de l'Industrie sur la montagne Sainte-Geneviève [à Paris]. Et j'ai aussi découvert le LDR, le Laboratoire de Dynamique des Réseaux, où j'ai rencontré Françoise Soulié-Fogelman, Gérard Weisbuch, et Maurice Milgram qui est devenu mon directeur de thèse. Et il y

avait une espèce de foisonnement d'idées, de gens qui s'intéressaient un peu à l'auto-organisation, qui est un peu le principe de base du deep learning et des réseaux de neurones.

Mehdi Khamassi [7.51]

C'est super intéressant, car tu évoques aussi des termes que je connais, de par ma formation [d'ingénieur], comme le filtrage adaptatif, mais dont je ne connaissais pas l'historique. Ainsi, il est vraiment fascinant de découvrir qu'il existe cette dimension de renommer certaines choses parfois pour avancer.

Que manque-t-il aux grands modèles de langage (LLMs) ?

Mehdi Khamassi [8.10]

Tu as justement évoqué une grande mode actuelle avec les grands modèles de langage, les «large language models» (LLMs), et l'IA générative, qui sont utilisés pour générer du contenu. Cela peut être du texte dans le cas des LLMs, mais également des images, des vidéos ou encore du son. En ce moment, on ne parle que de ça. Selon toi, l'attention mondiale qui leur est accordée est-elle justifiée ? Est-ce que c'est vraiment la voie royale de l'IA, celle par laquelle toutes ses ambitions, ses effets transformations, vont être réalisées ? Ou bien est-ce plutôt une voie latérale ? Dans ce cas, quelles sont les autres voies potentiellement transformatrices pour l'IA de demain ?

Yann LeCun [8.34]

Je pense que sur l'autoroute qui va nous mener vers des systèmes intelligents artificiels, les LLMs autorégressifs tels que nous les connaissons actuellement sont en quelque sorte une branche parallèle, voire une déviation, une diversion. Bien sûr, ils sont intéressants et nécessitent d'être développés, car ils présentent de nombreuses applications intéressantes et des choses à explorer. Ils sont très puissants pour certaines choses, mais ils ne suffisent pas.

Nous pourrions dresser une liste de caractéristiques qui définissent une entité comme étant intelligente, qu'elle soit naturelle ou artificielle. L'une de ces caractéristiques est la capacité à percevoir, par exemple, à percevoir l'état du monde. Cependant, les LLMs tels qu'ils existent actuellement ne sont pas capables de cela. Ils sont entraînés sur du texte et non avec des données provenant de capteurs du monde réel. Bien que cela puisse sembler une idée naturelle, c'est en réalité une tâche difficile à accomplir. Donc c'est la première chose : percevoir.

Ensuite, il y a la capacité de raisonner et de planifier. Malheureusement, les LLMs autorégressifs ne planifient pas. Leur raisonnement est extrêmement limité. On est pourtant un peu bluffé par ce qu'ils peuvent faire. C'est principalement parce qu'ils sont très bons dans la recherche d'informations et à leur adaptation à des problèmes connus. Cependant, ils montrent leurs limites dès qu'ils sont confrontés à des domaines en dehors de leur entraînement initial, qui est déjà vaste. Il y a des articles là-dessus de gens qui d'ailleurs viennent plutôt de l'IA classique, disons, qui montrent que les capacités de raisonnement des LLMs sont extrêmement limitées.

L'idée de raisonnement et celle de planification sont des caractéristiques essentielles du comportement intelligent. Cependant, les LLMs autorégressifs n'en sont pas capables. Ils ne comprennent pas le monde physique et ne pourront pas le faire sans changements radicaux de leur architecture. Ils ne

peuvent pas raisonner, ou en tout cas seulement de manière très simple. Ils ne peuvent pas planifier. Le raisonnement et la planification, c'est un petit peu la même chose, d'ailleurs. La planification est une sorte de raisonnement. Et surtout, les LLMs autorégressifs ne sont pas pilotables par des buts. Le comportement intelligent en général est piloté par des buts, des objectifs que le système tente d'atteindre. C'est ce que font les humains. C'est ce que font les systèmes de commande optimale, par exemple, dans lesquels il y a un but qui est réalisé par une fonction objective qui mesure si le but a été atteint. Et ensuite l'inférence procède par la recherche d'une séquence d'actions ou de sorties (de texte dans le cas d'un système de dialogue) qui va optimiser ce but, c'est-à-dire minimiser cette « fonction objectif ». On utilise ça en planification, en robotique, en commande optimale. Ça s'appelle *Model predictive control* en anglais, c'est une technique classique depuis 60 ans. Les LLMs ne savent pas faire ça. Et en fait, tout ça, ce sont des caractéristiques essentielles qu'on retrouve en biologie, chez tous les animaux et certainement les humains. Ceux-ci sont capables de planification vraiment très compliquée, beaucoup plus compliquée que ce qu'on sait faire avec les machines à l'heure actuelle.

Regardez un chat qui essaie de sauter sur un meuble, qui planifie sa séquence d'actions pour sauter de part en part dans la pièce. C'est une planification assez complexe. Et donc on manque d'ingrédients essentiels : ces capacités de planification, de satisfaction de buts. Et il manque aussi une autre chose qui permettrait ce raisonnement et cette satisfaction d'objectifs, et qui est la possession d'un modèle mental du monde, du monde physique ou de l'environnement dans lequel le système évolue. Les LLMs autorégressifs n'ont pas vraiment de modèle du monde. Ils en ont un, peut-être, qui est implicite.

Les psychologues séparent deux types de réactions chez les humains : ce qu'ils appellent « système 1 » et « système 2 ». C'est la nomenclature de Daniel Kahneman sur « thinking fast and slow » (la «pensée rapide» en opposition à la «pensée lente»). Le «système 1» englobe les actions subconscientes, pour lesquelles nous n'avons pas vraiment besoin de réfléchir, ni d'utiliser notre modèle [mental] du monde, ni de planifier. Ce système produit des réactions un petit peu rapides. Et le «système 2», en revanche, représente notre capacité à utiliser notre modèle [mental] du monde pour effectuer des tâches délibérément conscientes, et puis imaginer des scénarios et ensuite trouver une séquence d'actions qui va satisfaire les buts que nous nous sommes fixés. Les LLMs autorégressifs savent faire du «système 1», pas du «système 2».

Le défi de la recherche en IA pendant la décennie qui vient consiste à construire des systèmes capables de comprendre le monde, d'apprendre la physique intuitive, par exemple à la manière des bébés pendant leurs premiers mois de vie, qui ont la capacité de raisonner, de planifier et probablement de le faire de manière hiérarchique. En effet, on ne peut pas planifier des séquences d'actions complexes milliseconde par milliseconde. Ces systèmes devraient également être en mesure de satisfaire des objectifs.

Mehdi Khamassi [14.46]

D'ailleurs, chez les humains, il existe peut-être même plus que deux systèmes. Par exemple, en psychologie, des gens comme Olivier Houdé parlent de «système 3» pour décrire la coordination, la capacité à inhiber les systèmes non pertinents. Dans le contexte de la robotique, on parle de «méta-contrôleur». En somme, la hiérarchie est une notion constante, comme tu l'as mentionné. Donc au fond il y a toute cette question de la coordination [de systèmes].

Sur la question de savoir si les grands modèles de langage pourraient peut-être être intégrés à des choses ayant une connaissance plus physique du monde, j'ai l'impression que certaines équipes travaillent sur ça actuellement. Elles partent des LLMs et puis essaient de les connecter à autre chose.

À l'opposé, il y a des équipes qui adoptent une approche différente, et je dirais que c'est le cas pour nous [qui sommes dans le domaine de la robotique cognitive, appelée aussi robotique développementale]. Cette autre approche considère qu'il faut partir d'un apprentissage ancré dans le sensorimoteur, l'expérience et la mémoire, puis construire des modèles internes du monde, et que le langage va venir petit à petit.

Yann LeCun [15.29]

Oui.

Mehdi Khamassi [15.32]

Et peut-être que les deux approches sont intéressantes à développer, et aboutiront à des résultats assez différents.

Yann LeCun [15.36]

Oui.

Mehdi Khamassi [15.37]

Au cœur de tout cela, il y a la question cruciale que tu as soulevée : celle de la compréhension des représentations qui sont manipulées. Sans une compréhension de ce que signifie physiquement, par exemple, soulever un objet lourd, y associer une certaine sensation, ressentir le corps qui chauffe parce qu'on est en train de faire un effort, et parfois renoncer à cet effort en raison du risque de rupture pour le muscle. C'est en tout cas quelque chose qui existe dans le vivant.

D'accord avec Geoff Hinton que les LLMs comprennent ?

Mehdi Khamassi [16.00]

Je suis étonné d'avoir entendu récemment Geoff Hinton et Andrew Ng discuter et considérer que pour eux il y a un certain niveau de compréhension dans les grands modèles de langage. Est-ce que tu es en désaccord avec eux ?

Yann LeCun [16.13]

Non, je pense qu'il y a un certain niveau de compréhension, il n'y a pas de doute de ce côté-là, mais qui est quand même très ténu ou superficiel. Je ne pense pas qu'il y ait une réelle profondeur. En fait, j'ai même écrit un article de philosophie à ce sujet avec mon collègue Jacob Browning, dans la revue NOEMA, une revue philosophique disponible en ligne. Nous y mettons en avant le fait que la connaissance humaine dépasse de beaucoup celle qui est contenue dans la langue et dans le texte. C'est-à-dire que même si l'on prenait en compte la totalité des textes écrits depuis l'aube de l'humanité, cela ne représenterait qu'une partie, non pas infime mais assez petite, de la totalité des connaissances humaines. Le gros de la connaissance humaine, tout ce qu'on apprend dans la première année

de la vie par exemple, n'a absolument rien à voir avec le langage. C'est une connaissance du monde physique, qui est en d'un certain côté beaucoup plus complexe que la manipulation du langage.

Finalement, le langage, ce n'est pas si compliqué ! Après tout, le langage est un phénomène relativement récent dans l'histoire de l'humanité, ayant émergé il y a seulement quelques centaines de milliers d'années. Il est géré dans le cerveau par deux petites formations : l'aire de Wernicke et celle de Broca.

Mehdi Khamassi [17.28]

Avec un peu des fonctionnements distribués aussi.

Yann LeCun [17.31]

Bien sûr. Bien sûr. Tout le cerveau est activé quand on parle, bien sûr. Cependant, mais notre modèle du monde, c'est le cortex préfrontal. C'est vraiment la majeure partie du néocortex[1]. En réalité, cette fonction n'est absolument pas représentée ni implémentée par les LLMs autorégressifs.

Ce que je pense, et je prends un certain risque en disant ça, c'est que d'ici quelques années (c'est un peu difficile de savoir combien), nous pourrions avoir des systèmes capables de planification, de raisonnement, etc. On utilisera alors les modèles autorégressifs juste pour la traduction, la transcription, si on veut, d'idées abstraites en texte fluide. En effet, les systèmes autorégressifs présentent une qualité indéniable : les textes qu'ils génèrent sont vraiment bons, fluides et grammaticalement corrects, stylistiquement corrects. Donc il y a quand même des propriétés qu'il faut réutiliser.

Cependant, lorsqu'il s'agit de raisonnement, nous ne pensons pas nécessairement en termes de mots, mais en termes de représentations abstraites de ce que nous voulons dire ou des actions que nous voulons faire. Par conséquent, il faudra des systèmes qui soient basés là-dessus. J'ai fait quelques propositions dans ce sens, mais pour l'instant on n'a pas encore de système qui fonctionne bien et que je pourrais montrer.

À terme, je pense que les systèmes autorégressifs finiront par disparaître. En effet, ils ne sont ni pilotables ni contrôlables, car ils ne sont pas conçus pour remplir des objectifs. Ils racontent des bêtises.

Mehdi Khamassi [19.20]

On va y revenir d'ailleurs.

(rires)

Yann LeCun [19.22]

C'est ce qu'on appelle les « hallucinations », que nous appelions auparavant des « fabulations » en fait, etc. Je pense donc que ces systèmes vont finir par disparaître.

---

1. [NdA] En neurosciences, la question des circuits cérébraux impliqués dans l'apprentissage de modèle(s) interne(s) (mental/mentaux) du monde est très débattue et encore à défricher. Un modèle du monde, ou encore modèle interne, est quelque chose qu'on apprend au fil de notre expérience sensorimotrice avec le monde qui nous entoure, et qui nous permet de prédire l'effet de nos actions. C'est grâce à ça qu'on reconnait les situations familières, dans lesquelles on prédit bien, et donc dans lesquelles on n'a pas besoin de changer notre comportement, des situations nouvelles ou surprenantes, qui indiquent qu'on a besoin d'analyser la situation pour déterminer si un nouveau comportement est nécessaire. Par exemple, quand on met le pied sur un tapis roulant à l'arrêt, on éprouve une sensation déstabilisante liée à une erreur de prédiction (i.e., on s'attendait à du mouvement qui ne se produit pas), qui nous oblige à changer de posture pour ne pas tomber. Dans le cerveau, on considère que le cortex préfrontal joue un rôle important, mais aussi d'autres structures telles que le cervelet, l'hippocampe et d'autres parties du cortex selon les modalités sensorielles et le contexte social ou non social de l'action.

Ce qui m'a surpris chez Geoff Hinton, pour revenir directement à la question, c'est qu'il a eu une espèce de révélation il y a quelques mois en jouant un petit peu avec les LLMs disponibles chez Google et ailleurs. Sa révélation correspond au fait que toute sa vie il a cherché, c'est en quelque sorte la quête de sa vie : une procédure d'apprentissage qui fonctionne aussi bien que ce qu'on observe dans le cerveau. C'est ça qu'il a travaillé sur les machines de Boltzmann, la rétropropagation, les algorithmes de recirculation, les RBM et autres. Il y a eu toute une série de travaux. Et sa révélation a consisté à dire : « il est possible que la rétropropagation, en fait, soit la réponse ». Ce n'est pas la réponse sur ce que fait le cerveau, mais ça marche au moins aussi bien, sinon mieux. Et la raison pour laquelle je dis que ça marche mieux, c'est qu'on voit les capacités de ces réseaux de neurones, qui ne sont en fait pas très grands par rapport à la taille du cerveau, et qui ont ces capacités assez incroyables, un peu supérieures au cerveau. Par exemple, on peut maintenant créer des systèmes de traduction qui traduisent des centaines de langues, et des systèmes de reconnaissance de la parole qui reconnaissent des milliers de langues. C'est tout incroyable ! C'est surhumain ! Et avec des réseaux qui sont en fait pas si gros.

Donc il [Geoff Hinton] s'est dit que peut-être un neurone artificiel [a déjà une très grande capacité de calcul]. Ceci ne correspond pas à l'argument évoqué depuis très longtemps en neurosciences, selon lequel l'intelligence artificielle est très très simplifiée par rapport aux neurones réels [aux neurones biologiques], et selon lequel il faudrait plusieurs centaines de neurones artificiels pour accomplir la fonction équivalente à un neurone biologique. Cependant, la constatation de Geoff, c'est le contraire : « en fait, non, pas du tout, un neurone artificiel peut en réalité remplir la fonction de plusieurs neurones biologiques. Peut-être que la biologie a besoin de davantage de neurones à cause de la séparation entre synapses excitatrices et inhibitrices[2], et parce qu'il faut tout un circuit pour essayer de calculer des gradients équivalents à la rétropropagation, mais on ne peut pas le faire avec les mêmes neurones que ceux utilisés pour la rétropropagation. En somme, il en est venu à penser que nous sommes plus proches de l'intelligence humaine que nous ne le pensions. Cependant, pour lui, le problème principal reste de trouver la procédure d'apprentissage. De mon côté, j'ai toujours pensé que la rétropropagation était un petit peu la meilleure manière, la plus efficace, de calculer des gradients, et je ne voyais pas d'autres manières pour l'apprentissage que par estimation de gradients. Donc, s'il nous faut la rétropropagation, c'est bon, et ce sur quoi il nous faut travailler, ce sont plutôt les paradigmes d'apprentissage, tels que l'autosupervision, la gestion de la certitude, le raisonnement, et autres.

Des risques existentiels pour l'humanité ?

Yann LeCun [22.31]

Donc il [Geoff Hinton] s'est dit : « l'obstacle que je percevais comme le plus important, l'apprentissage, a disparu. Donc on est à deux pas de l'intelligence de niveau humain. » Donc du coup ça vaut le coup de réfléchir aux risques, chose qu'il n'avait jamais envisagée auparavant. Il a donc eu la réaction immédiate, de premier ordre, que tout le monde a quand on réfléchit à ça [aux risques de l'IA] pour la première fois : c'est le scénario à la Terminator, auquel je ne crois absolument pas.

(rires)

2. [NdA] Alors qu'un neurone artificiel peut avoir à la fois des synapses excitatrices (c'est-à-dire dont le poids synaptique est positif) et des synapses inhibitrices (dont le poids synaptique est négatif), car mathématiquement on peut combiner des choses positives et des choses négatives au niveau du même neurone. À l'inverse, le système nerveux utilise des éléments chimiques distincts pour mettre en œuvre des synapses inhibitrices (e.g., le GABA) et des synapses excitatrices (e.g., le Glutamate), et un neurone biologique donné ne peut pas émettre à la fois du GABA et du Glutamate.

Donc on n'est pas d'accord là-dessus. Pas du tout. D'abord, je ne crois pas que les LLMs nous mèneront à l'intelligence de niveau humain. Je pense qu'il y a encore beaucoup de sauts conceptuels à faire avant ça. Et puis, deuxièmement, je ne crois pas non plus que les histoires de machines super intelligentes causeront l'extinction de l'humanité, ou même un danger en fait.

Mehdi Khamassi [23.18]

Tu as fait une réponse très riche. Je vois plein d'éléments.

Yann LeCun [23.21]
Plein plein de choses !

(rires)

Deux liens avec les neurosciences

Mehdi Khamassi [23.23]

Mais on va en discuter, on va approfondir. J'ai envie de faire deux liens avec les neurosciences dans ce que tu dis. (1) Du fait que, finalement, même la rétropagation du gradient a toujours été considérée comme quelque chose qui n'est pas plausible du point de vue qui se passe dans le cerveau, mais en fait il y a de plus en plus d'équipes [de recherche] qui disent : « si, si, regardez, on peut le faire d'une certaine manière, avoir un autre réseau parallèle qui apprend des choses et qui les retransforme dans l'autre direction. » Il y a des gens qui travaillent là-dessus, donc pour moi ce n'est pas complètement non plausible. (2) Et puis l'autre aspect concerne le langage : dans ma compréhension de certains processus de production, par exemple de séquences de mots – et je ne suis pas un spécialiste de langage, je travaille plutôt du côté décision et planification –, mais il semble que le cerveau réutilise, recycle justement, des liens anatomiques entre le cortex préfrontal et les ganglions de la base pour le séquencement des mots comme il l'a fait pour le séquencement des actions. Donc il y a cette idée, un peu développementale, qu'il y a des choses qu'on apprend pour interagir avec le monde, construire des modèles, planifier, et qu'ensuite on peut réutiliser ça pour le langage. Je trouve que ça va vraiment dans le sens de ce que tu dis.

Apprentissage autosupervisé

Yann LeCun [24.16]

Oui. En fait, la raison pour laquelle je disais précédemment que le langage est finalement simple pour ce qui est de l'intelligence artificielle, c'est que le langage manipule des symboles discrets, les mots, qui sont en nombre fini, et on sait que depuis quelques années l'apprentissage autosupervisé est devenu complètement dominant. Tous les systèmes de vision aussi bien que de langue aujourd'hui, de reconnaissance de la parole, etc., sont préentraînés de manière autosupervisée.

Alors, qu'est-ce que ça veut dire l'entraînement autosupervisé ? Ça veut dire ne pas entraîner un système pour une tâche particulière, mais lui apprendre à représenter les données relativement

indépendamment de la tâche, disons. Et certains des succès [ont été obtenus grâce à ce type d'entraînement autosupervisé], qu'on peut essayer de résumer en un concept très simple : si on prend une entrée, que ce soit un texte, une vidéo, une image, ou même une paire d'images, et si on corrompt cette entrée d'une manière ou d'une autre – on peut la corrompre par masquage, c'est-à-dire, par exemple, que dans un texte on remplace certains mots par un marqueur BLANC, ou dans une image on fait une corruption de l'image, c'est-à-dire qu'on fait une distorsion, on la floute, on change les couleurs, etc., et puis dans une vidéo on peut aussi masquer une partie de la vidéo – et ensuite on entraîne un très gros système d'apprentissage (e.g., un réseau de neurones, une architecture de *deep learning*) à remplir les trous, c'est-à-dire à prédire ce qui manque.

Donc si on fait ça avec du texte, on retire typiquement 10 à 20% des mots, et puis on entraîne un gros réseau de neurones à prédire les mots qui manquent. Se faisant, le système élabore une représentation interne du texte qui contient toute l'information sur la sémantique, la grammaire, et tout. C'est ça qui a conduit à une révolution en NLP (*Natural Language Processing* ; en français : Traitement Automatique du Langage), en paroles, ces derniers 5 ans en gros, 5-6 ans à peu près.

Mais on ne peut jamais faire cette prédiction exactement, car on ne peut jamais prédire exactement quel mot se substitue au blanc. Si je dis « le chat chasse la <blanc> dans la cuisine », ça peut être une souris, mais ça peut être la lumière d'un laser, ça peut être, etc.

Mehdi Khamassi [26.36]

Ça dépend d'un contexte.

Yann LeCun [26.37]

Voilà ! Ça dépend d'un contexte qu'on n'a pas forcément. Il y a toujours une incertitude dans la prédiction, et pour pouvoir faire de l'apprentissage autosupervisé si on le fait de manière générative, c'est-à-dire en reconstruisant ce qui manque, il va falloir pouvoir gérer l'incertitude dans la prédiction. Alors si c'est du texte, c'est facile, parce qu'il n'y a qu'un nombre fini de mots dans le dictionnaire, donc il suffit de produire une distribution de probabilités sur tous les mots du dictionnaire. C'est un grand vecteur de nombres entre 0 et 1 avec somme à 1. Pas compliqué, on sait le faire.

Abandonner les modèles génératifs pour l'image et la vidéo

Yann LeCun [27.09]

Mais par contre si c'est une image, par exemple, ou un segment de vidéo, on ne sait pas le faire ; on ne sait pas bien représenter des distributions de probabilités sur l'ensemble de toutes les images ou de toutes les vidéos. La raison en est que c'est continu, que c'est de haute dimension, etc. Et c'est même pire que ça ! On ne sait même pas correctement faire de la reconstruction dans ce contexte, parce que si on prend le segment initial d'une vidéo et qu'on demande au système de prédire ce qui va se passer par la suite, il y a plein, plein, plein de futurs possibles ! Ou plausibles ! Et donc si on demande au système de faire une prédiction, la seule chose qu'il peut faire est de prédire une espèce de moyenne de tous les futurs possibles, ce qui est une image floue ou une vidéo floue. Donc il faut

des systèmes à variables latentes[3], ou qui puissent représenter des prédictions multiples, et c'est extrêmement compliqué.

Une des meilleures condamnations récentes a été : si on veut entraîner le système avec des informations sensorielles telles que la vidéo, et le faire apprendre comment fonctionne le monde en le regardant au passé, un peu comme font les bébés, il faut abandonner l'idée de modèle génératif[4], et se baser sur des modèles qui élaborent des représentations internes du monde, qui éliminent les détails qui ne sont pas prédictibles, et qui font la prédiction dans l'espace de représentations abstraites. J'appelle ça les JEPA, les *Joint Embedding Predictive Architectures*[5]. La raison pour laquelle je suis arrivé à ça, c'est que toutes les méthodes d'autosupervision, enfin d'entraînement par autosupervision en image, qui marchent bien sont toutes des méthodes de ce qu'on appelle *joint embedding*. C'est-à-dire que ce sont des méthodes dans lesquelles on n'essaie pas de reconstruire, mais si on a par exemple une image distordue et qu'on essaie d'entraîner un système à dire « la représentation doit être la même parce que c'est le même contenu », on ne cherche pas à reconstruire l'image originale ou même l'image distordue, on essaie simplement de prédire la représentation interne de l'une à partir de l'autre. Ça veut dire abandonner l'idée de modèles génératifs, qui est le truc le plus à la mode aujourd'hui.

(rires)

Donc ce n'est pas une idée qui est facile à vendre. Mais on travaille dessus. Donc déjà pour nous, la révolution des LLMs autorégressifs date de 2 ans, ou 3, et depuis 2 ans ou 3 ans on travaille sur comment lever les limitations de ces systèmes.

Deuxième partie de l'audition : Apprendre des modèles internes du monde

Mehdi Khamassi [0.08]

Alors, tu as mentionné une des pistes, qui est de creuser comment apprendre des modèles génératifs du monde [en fait non pas génératifs mais plutôt : des modèles internes du monde], et ainsi notamment d'apprendre dans l'interaction physique avec le monde quelles sont les conséquences de nos actions, qu'est-ce que ça produit comme effet sur le monde. Dans un papier d'opinion de l'année

---

3. [NdA] En physique, une variable *latente* représente un événement ou une cause caché(e), que les calculs doivent tenter d'inférer pour pouvoir expliquer de manière plausible (de manière fortement probable) la succession d'événements ou variables que l'on a observé(e)s, donc une succession de données qu'un expérimentateur a mesurées.

4. [NdA] On appelle *modèles génératifs* des modèles qui essaient de reconstruire l'information qui manque dans les données. Pour cela, ces modèles doivent par apprentissage avoir réussi à retrouver le processus qui par une certaine régularité génère les données observées. Par exemple, à un péage autoroutier on pourrait faire apprendre à un modèle génératif qu'à chaque fois qu'on détecte qu'un véhicule passe, avec en plus un capteur au sol qui mesure le poids du véhicule, on observe juste après le véhicule passer sur une vidéo de surveillance. Alors, même si parfois il manque des données vidéos (par exemple en cas de panne momentanée, ou de problème de connexion réseau à la caméra), le modèle pourra reconstruire approximativement une vidéo de moto qui passe, de voiture qui passe, de camion qui passe, ou des moments où rien ne passe, ceci en fonction du poids mesuré à chaque instant : moins de 200 kg, moins d'une tonne, plusieurs tonnes, ou un poids à peu près nul, respectivement à ces quatre cas. Mais dans des situations moins contrôlées, il est quasiment impossible d'apprendre un modèle génératif, car le nombre de possibilités est trop grand, voire infini, et on n'a pas assez de capteurs qui nous permettraient de bien prédire ce qu'on va observer dans la vidéo.

5 En français, on écrirait que les JEPA sont des architectures algorithmiques d'apprentissage pour prédire des représentations d'images (typiquement), et en particulier des représentations compatibles pour des images similaires.

dernière [2022], je crois, tu as proposé toute une architecture. Et une chose que j'ai notée, c'est que tu proposes qu'il faudrait [faire apprendre à un agen] un seul grand modèle génératif. Ça soulève plein de questions, parce qu'il y a aussi des défauts d'avoir plein de modèles différents. Donc tu dis effectivement qu'avoir un modèle spécifique à chaque tâche n'est pas forcément efficient d'un point de vue computationnel. On pourrait se dire qu'il y a néanmoins la question de [comment construire des modèles sur] plusieurs échelles, parce que parfois il faut raisonner avec des choses qui vont se succéder à l'échelle de quelques secondes, mais parfois il y a des choses qui ont des conséquences à de plus longs délais, ou à une échelle spatiale différente. Au fond, comment faut-il faire pour gérer ces différentes échelles, ces différentes sous-tâches ou *skills* (compétences en français) qui pourraient être appris en lien avec un modèle ?

Yann LeCun [1.03]

Alors deux ou trois choses. La première chose que je voudrais corriger, c'est que ce n'est pas un modèle génératif.

Mehdi Khamassi [1.08]

Pardon, pardon, je voulais dire un modèle interne du monde. J'ai dit génératif ? Oui bien sûr.

Yann LeCun [1.12]

Oui voilà. Non génératif dans l'espace d'entrée, mais effectivement on pourrait dire génératif dans un espace de représentations. Toutes les prédictions se passent dans des espaces de représentations. Donc il y a une version de cette architecture, que je l'appelle JEPA [Joint Embedding Predictive Architecture], qui peut être entraînée par autosupervision, simplement en observant le monde. Et puis si on veut un modèle causal, c'est-à-dire qu'li puisse prédire les conséquences des actions de l'agent, il faut que l'agent puisse effectivement faire une action sur le monde, observer le résultat, et ensuite entraîner son modèle à prédire le résultat. Avoir un modèle de ce type-là permettrait de faire de la planification. Et ça, ça rentre complètement dans le cadre classique de la commande optimale, le *Model-Predictive Control* (c'est-à-dire le contrôle de l'action en fonction des conséquences prédites par un modèle). Mais ce que je propose après, c'est une version hiérarchique de ça, c'est à dire dans lequel les représentations ne sont pas à un niveau mais à plusieurs niveaux, et dans lequel la prédiction à un haut niveau d'abstraction permet de faire de la prédiction à plus long terme.

Par exemple, si je planifie d'aller de New York à Paris, je suis en fait à Paris, là, aujourd'hui *(rires)*, mais si je prévois d'aller de New York à Paris, la première chose que je fais c'est de décomposer ça en deux sous-tâches : d'abord attraper un taxi pour aller à l'aéroport, c'est la meilleure manière de faire à New York, malheureusement ; et puis ensuite prendre un avion pour Paris. Donc maintenant j'ai une première sous-tâche qui est : aller à l'aéroport. Pour aller à l'aéroport, il me faut attraper un taxi. Pour attraper un taxi, il faut que je descende dans la rue, et que j'appelle un taxi. Comment je descends dans la rue ? Il faut que je me lève de ma chaise, que j'aille vers l'ascenseur ou l'escalier, que je descende, etc. Comment je me lève de ma chaise ? Etc. Donc on peut tout décomposer jusqu'à ce qu'on arrive à des actions élémentaires de contrôle musculaire, qui doivent être planifiées milliseconde par milliseconde. Mais on ne peut pas faire la planification au niveau le plus bas du contrôle musculaire de toute la trajectoire entre être assis dans son bureau à New York et arriver à Paris. Ce n'est pas possible. Donc il faut nécessairement faire cette décomposition hiérarchique.

Alors la question est : combien de cette décomposition hiérarchique peut-on contenir dans notre cortex préfrontal, dans notre modèle du monde ? Est-ce qu'il faut faire cette décomposition hiérarchique avec un modèle hiérarchique ? Ou bien est-ce qu'au contraire il faut simplement un modèle un petit peu plus simple mais qui est configurable à la situation qu'on considère ?

Un seul ou plusieurs modèles internes du monde ?

Yann LeCun [3.45]

Donc là j'en viens un peu à répondre à la question qui est : il y a un avantage à avoir un moteur de modèle du monde unique – en tout cas dans le cas d'un cortex préfrontal, mais pour une machine on ne sait pas encore –, et qui serait configurable pour la tâche à accomplir. Et la raison d'avoir un modèle unique, ou disons en petit nombre, est qu'on fait ainsi des économies d'échelle, c'est-à-dire que toutes les situations qu'on rencontre ont toutes quelque chose en commun : elles se déroulent dans le monde réel, ou elles ont à voir avec des interactions avec des humains, ou d'autres agents, etc. Donc notre modèle du monde a une certaine logique interne, due à la logique interne du monde.

Il serait donc probablement inefficace d'avoir un modèle complètement séparé dans notre tête pour toutes les situations qu'on a à rencontrer, par opposition à un modèle unique qui serait configurable et qui pourrait réutiliser les composantes et les propriétés qui sont communes à tous les problèmes auxquels on doit faire face. La logique est toujours la logique, etc. Le monde tridimensionnel est toujours tridimensionnel. La physique intuitive est toujours la physique intuitive. Bon, il y a des variations, bien sûr, qui sont un peu déroutantes, par exemple quand on est dans l'espace en gravité 0, ou dans un jeu vidéo dans lequel la physique est complètement changée. Mais on peut s'adapter à ça. Et il y a un petit peu d'indices qui montrent que les humains ont en fait une sorte de modèle unique. Et cet indice est le fait qu'on ne peut résoudre qu'une tâche consciente à la fois. Plus précisément, quand on se focalise sur une tâche, on ne peut résoudre que cette tâche, à l'exception de toutes les autres. On peut faire plusieurs tâches subconscientes à la fois, mais une seule tâche consciente. Donc ça tendrait à suggérer qu'on a un modèle du monde qui nous permet de prédire les conséquences de nos actions et qui nous permet de planifier.

Mehdi Khamassi [5.54]

C'est une question intéressante parce que c'est vrai qu'il y a des avantages et des inconvénients dans chaque sens. En particulier, si on a plein de modèles séparés, comment on généralise ?

Yann LeCun [6.01]

Oui.

Mehdi Khamassi [6.02]

Donc s'il y a un modèle qui contient quelque chose du ressort de la physique intuitive et qu'on ne l'a pas dans l'autre [modèle], c'est dommage, car il faut le réapprendre. Après on peut imaginer quand même que dans des périodes *offline* (hors ligne, en français), pendant le rêve ou autre, on réas-

semble des choses, on trouve des points communs. Et peut-être que ça irait dans le sens de construire aussi une représentation unique. Et puis à l'inverse, il y a quand même aussi des coûts, il me semble, des conséquences telles que par exemple des interférences, quand on a un seul modèle. Ceci peut faire que quelque chose qui se passe à une autre échelle ou pour une autre tâche peut venir interférer avec la décision qu'on va prendre là.

Yann LeCun [6.29]

Oui.

Mehdi Khamassi [6.30]

Et ça peut dans certains cas ne pas être approprié. Comment tu proposes de gérer ça ?

Yann LeCun [6.35]

Alors je ne sais pas. Dans ma proposition, dans l'article que tu mentionnes, et dont le titre est « A path towards autonomous machine intelligence » – alors je ne l'appelle plus « autonomous machine intelligence » (en français, intelligence des machines autonomes), maintenant, parce que ça fait peur aux gens quand on dit « intelligence autonome ». J'appelle ça « objective-driven AI » (en français, l'intelligence artificielle dirigée vers des objectifs). C'est l'idée que le comportement du système est piloté par des objectifs, que le système doit optimiser des fonctions objectives qu'il doit minimiser à travers ses actions. Et ça rend le système non seulement pilotable mais aussi plus sécurisé, c'est-à-dire qu'on peut mettre dans ses fonctions d'objectifs des rambardes qui empêchent que le système fasse des choses un peu folles qui sont contraires à l'intérêt des utilisateurs, par exemple. On est très familier de ce genre de choses pour la commande optimale, par exemple : on met des limites à la poussée d'un moteur-fusée, ou des choses comme ça, qui sont des limites dures, des contraintes sur les commandes qu'on peut faire pour éviter les désastres. Donc c'est un petit peu la même chose. Et je ne donne pas dans cet article de solution sur comment dynamiquement configurer le modèle du monde pour l'adapter à la tâche considérée.

J'ai un module en haut, qui s'appelle « le configurateur », qui est censé configurer la fonction du modèle du monde. Maintenant on a des méthodes qui sont quand même bien, par exemple avec les *Transformers*[6], on a une architecture qui est configurable intrinsèquement. C'est-à-dire qu'on peut prendre une architecture de *Transformers*, séparer son entrée – donc une série de vecteurs – en deux sections, utiliser une section comme étant la vraie entrée, et l'autre section comme étant des variables de configuration qui pilotent et qui changent la relation entrée/sortie du reste du réseau. Donc ça vous donne un outil assez simple. Il y a plusieurs manières de faire ça. Il y a une autre idée qu'on appelle les hyper réseaux, *hyper networks*, mais qui est beaucoup moins efficace par certains côtés. Donc ça nous donne un outil, c'est ce que je recommande dans l'article, pour utiliser des architectures à la *Transformers*, des choses un petit peu similaires, pour le prédicteur, c'est-à-dire le modèle prédictif.

---

6. Les *Transformers* sont une architecture algorithmique qui utilise des mécanismes d'attention pour établir des dépendances globales entre les entrées et les sorties. Ces mécanismes d'attention permettent notamment de mettre l'accent sur des dépendances entre deux éléments distants (*i.e.,* non directement consécutifs) au sein d'une séquence temporelle d'éléments (par exemple entre le 1er et le 5^{e} élément). Les *Transformers* permettent du coup une plus grande parallélisation des calculs puisque l'on n'est plus contraint de traiter séquentiellement les éléments.

Une motivation intrinsèque à apprendre

Mehdi Khamassi [9.07]

Un deuxième ingrédient que tu proposes dans cet article, et qui sont des choses qui sont très étudiées en robotique développementale, justement le côté apprentissage ouvert, tout au long de la vie, et le fait de réutiliser des connaissances apprises dans un contexte vers un autre contexte, avec des inspirations de la psychologie cognitive, bien sûr, cette notion de motivation intrinsèque.

Yann LeCun [9.28]

Oui.

Mehdi Khamassi [9.29]

L'idée de développer des motivations initiales du système pour qu'il puisse être curieux d'acquérir des connaissances, et se définir ensuite lui-même des sous-buts pour pouvoir éventuellement satisfaire ses motivations. Mais c'est une question très difficile : comment définir des sous-buts pertinents ? Alors qu'est-ce que tu proposes dans ce sens-là ?

Yann LeCun [9.48]

Oui, alors c'est pareil, je n'ai pas de réponse très claire, si ce n'est des exemples qui montrent que certains critères, certains objectifs de ce type-là, dans certains contextes, conduisent le système à apprendre des représentations qui sont appropriées. Mais bon, le gros avantage des objectifs intrinsèques, c'est qu'à la différence d'objectifs externes qu'utilisent les gens qui travaillent en apprentissage par renforcement par exemple, qui sont des objectifs inconnus de l'agent et qu'il doit interroger en prenant une action, un objectif intrinsèque est une fonction qui est calculée par l'agent lui-même.

Dans l'apprentissage par renforcement, il y a un objectif, mais l'objectif est inconnu de l'agent. La seule manière pour l'agent de connaître l'objectif, c'est de prendre une action dans le monde et d'attendre que le monde lui donne une récompense ou une punition, c'est-à-dire un scalaire qui lui dit que la réponse était bonne ou pas bonne. Et ça, c'est extrêmement inefficace[7]. C'est ce qui m'a conduit

7. C'est inefficace pour deux raisons qui sont bien connues dans la recherche sur l'apprentissage par renforcement : (1) d'une part, il faut que l'humain ait défini au préalable la fonction de récompense qui va guider l'apprentissage de l'agent artificiel, c'est-à-dire une fonction qui détermine quelles actions sont bonnes et quelles actions ne le sont pas. L'agent ne peut pas l'apprendre tout seul sans qu'une fonction de récompense ne lui soit donnée. On donne parfois le contre-argument partiel comme quoi dans le cerveau biologique, il semble bien y avoir des mécanismes qui ont été sélectionnés au cours de l'Évolution et qui permettent de considérer certains événements comme automatiquement récompensants (*e.g.,* le fait de trouver de la nourriture, le fait de trouver un abri, le fait de voir un individu sourire, etc.), et d'autres événements comme « punissants », c'est-à-dire associés à des récompenses négatives (*e.g.,* se retrouver nez à nez avec un prédateur, risquer de tomber d'une falaise, etc.). Mais cela n'est pas suffisant pour produire des récompenses dans toutes les situations où il y a matière à apprendre (*e.g.,* trouver la solution à un casse-tête ne rentre dans aucun de ces cas stéréotypiques, et pourtant cela semble bien générer un sentiment de satisfaction, comme si notre cerveau produisait un signal de récompense intrinsèque, c'est-à-dire généré en interne, contrairement aux exemples de récompenses extrinsèques qui précèdent qui peuvent venir de l'extérieur de l'agent) ; (2) même si l'on définit à l'avance un certain nombre d'événements comme étant des récompenses pour l'agent, ces événements sont trop rares au milieu d'une écrasante majorité d'événements « neutres », *i.e.,* non récompensants, pour que l'apprentissage puisse progresser suffisamment vite. Par exemple, si l'on décide de donner une récompense à un robot lorsqu'il aura réussi à monter un meuble, il y a un tel nombre d'actions intermédiaires à réaliser et qui aboutissent à des événements neutres (*e.g.,* enfoncer une vis jusqu'au bout ne donne pas lieu à une récompense), au milieu d'un nombre encore plus grand d'actions non pertinentes

à dire que l'apprentissage par renforcement ne peut être que la cerise sur le gâteau, parce que c'est tellement inefficace au niveau de la quantité d'essais et d'erreurs, de la quantité d'exemples, que ça ne peut pas représenter le principal de l'apprentissage dans le monde réel. Ça marche néanmoins bien pour les jeux parce qu'on peut jouer des millions de parties en quelques heures. Mais ça ne marche pas dans un environnemental réel.

Donc, à la différence d'objectifs inconnus, qu'on doit interroger en prenant une action, un objectif intrinsèque est une fonction qui est calculée par l'agent lui-même. Donc c'est un autre réseau de neurones, qui doit être différentiable, donc on peut rétropropager des gradients[8] à travers, donc on peut essayer de prendre des actions qui optimisent ces objectifs. On peut par exemple les utiliser pour rétropropager des gradients pour entraîner des modèles de perception, de prédiction, etc. Donc c'est beaucoup plus efficace si on peut conduire un système à apprendre des concepts à partir de motivations intrinsèques.

Bien sûr, je suis loin d'être le premier à proposer ça. Pierre-Yves Oudeyer travaille là-dessus depuis très longtemps, et pas mal d'autres gens ont proposé ce genre de choses aussi. Et puis en commande optimale c'est un petit peu la même chose. Donc ce n'est pas une idée nouvelle, mais disons que la présenter comme étant une espèce d'ingrédient essentiel par opposition à l'apprentissage par renforcement, je pense que c'est quelque chose sur lequel il faut se focaliser. En effet, ça permet une forme d'apprentissage auto supervisé.

Je vais juste donner un exemple : on a travaillé sur une méthode, récemment, dont je parle un peu dans l'article, qui s'appelle VICREG. Ça veut dire : Variance, Invariance, Covariance, Régularisation, en anglais. Cette méthode consiste à entraîner une de ces architectures de *Joint Embedding*, dans laquelle on prend une image et une image distordue, par exemple, on les fait tourner toutes les deux, on les passe par deux copies d'un réseau de neurones – mais ils n'ont pas besoin d'être identiques ; ils peuvent être différents. Et en fait ça peut être une image, de l'audio, ça peut être des choses différentes, peu importe. Mais dans notre exemple, considérons un cas simple : des images, des vues différentes de la même scène, par exemple. Et ce qu'on veut, c'est, du fait que le contenu est identique, que les représentations soient aussi identiques. Donc on peut entraîner le système à produire des représentations identiques pour ces pairs d'images qu'on lui donne. Malheureusement, si on ne fait que ça, le système collapse. Il dégénère. Il produit des représentations qui sont constantes, c'est-à-dire qu'il ignore complètement les entrées, [au lieu de produire des représentations qui dépendent des entrées du réseau,] et comme ça les représentations sont toujours identiques. Donc il faut un moyen d'éviter ce *collapse*, ce type d'échecs. Il y a deux classes de méthodes pour ça : (1) une classe de méthodes qu'on appelle *les méthodes contrastives*. Cela consiste à montrer aussi [en plus] des pairs d'images qui sont différentes, et à repousser ainsi les représentations calculées les unes des autres durant l'apprentissage. En fait, ça, c'est une très vieille idée que j'avais proposé en 1993, dans le contexte de ce qu'on appelait *les réseaux siamois*, les réseaux identiques, pour apprendre des *embeddings*, *i.e.,* des représentations d'images, qu'on puisse comparer par la suite. Et puis cette idée a été un petit peu ravivée par un article dont Geoff Hinton est co-auteur, qui s'appelle SimCLR[9], donc aussi une méthode contrastive. Et puis il y en a d'autres : CPC, etc. Mais en fait je suis devenu

qui aboutissent aussi à des événements neutres (*e.g.,* dévisser une vis), que le robot n'aurait statistiquement aucune chance de trouver tout seul la solution.

8. Un gradient est un différentiel entre une sortie souhaitée et une sortie réellement produite par un réseau de neurones, donc un signal d'erreur qui peut être utilisé pour guider l'apprentissage en corrigeant les paramètres du réseau, c'est-à-dire les « poids synaptiques » qui connectent les « neurones » du réseau. Or les réseaux de neurones profonds ont plusieurs couches de neurones, qui effectuent des traitements successifs sur l'information fournie en entrée pour aboutir à une réponse en sortie. Pour permettre à l'apprentissage d'adapter les paramètres des couches intermédiaires, il faut donc propager le signal d'erreur, le gradient, de la couche de sortie du réseau vers les couches intermédiaires, donc le « rétropropager ».

9. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

assez négatif sur ces méthodes contrastives parce qu'elles ne marchent pas très bien en hautes dimensions. Plus précisément, si les représentations sont de hautes dimensions il faut beaucoup d'exemples contrastifs pour arriver à repousser toutes les représentations les unes des autres. (2) Donc cette idée de VICREG est différente. Elle consiste à essayer de maximiser le contenu informationnel de la représentation on s'assurant que chacune des variables de la représentation n'est pas constante, ceci en mettant un critère sur sa variance sur un certain nombre d'exemples, et puis aussi en décorrélant des pairs de variables de manière à ce que le contenu informationnel du vecteur soit un peu maximisé. C'est un peu douteux au niveau mathématique parce que ça consiste à pousser vers le haut une estimation du contenu informationnel qui est en fait une borne supérieure, et pas inférieure. *(rires)* Mais bon ça marche à peu près bien. *(rires)* Donc on est assez content de cette méthode. Il y a aussi d'autres méthodes différentes pour empêcher le *collapse*. Par exemple, il y a une méthode qu'on a proposée récemment et qui s'appelle I-JEPA[10] : Images JEPA. Mais bon, ça, c'est une méthode de masquage. Et puis il y en a quelques autres proposées par des collègues ici à Paris, dont une qui s'appelle DINO, DINOV2[11] en particulier. C'est une autre méthode encore pour empêcher le *collapse*. Il y en a une bonne douzaine comme ça. Mais je pense que c'est dans cette direction-là qu'il faut se diriger.

Optimisation multi-objectifs

Mehdi Khamassi [15.48]

Dans ce que tu dis, il y a eu aussi quelque chose d'un peu complémentaire dans des directions à creuser : quelque part, quand je t'entends, j'entends une opposition entre apprentissage par renforcement et motivation intrinsèque. Alors qu'on peut se dire que [ça peut être tout à fait complémentaire] : du côté de l'apprentissage par renforcement il y a une mécanique – l'apprentissage par erreur de prédiction, qui d'ailleurs, si on la transforme sous forme probabiliste, commence à devenir comme de l'inférence bayésienne[12], de l'inférence active[13].

Yann LeCun [16.08]

Oui.

Mehdi Khamassi [16.09]

Ce n'est finalement pas si différent. Et puis de l'autre côté, il y a la notion qu'on peut combiner plusieurs objectifs, qu'on peut être multi-objectifs.

---

10. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., ... & Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15619-15629).
11. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
12. L'inférence bayésienne est une méthode statistique qui consiste à calculer la probabilité d'événements en fonction d'observations, de connaissances a priori, et du niveau d'incertitude lié à ces éléments. Ce calcul repose essentiellement sur l'application du théorème de Bayes.
13. L'inférence active est un cadre conceptuel et computationnel qui décrit comme un agent peut apprendre à se représenter son environnement de manière active, en cherchant à prédire les propriétés et événements de cet environnement, et en apprenant sur la base d'erreurs de prédiction. Dans ce cadre, l'un des moteurs principaux de l'action est la quête d'informations permettant de réduire l'incertitude dans cette représentation (i.e., valeur épistémique). Ce cadre théorique, promu par Karl Friston, chercheur en neurosciences computationnelles, est lié aux formalismes de la physique statistique et de la théorie de l'information.

Yann LeCun [16.14]

Tout à fait.

Mehdi Khamassi [16.15]

Avec des objectifs extrinsèques, *e.g.,* des fois j'acquiers de l'énergie quand je mange de la nourriture ; et des objectifs intrinsèques, *e.g.,* des fois j'acquière de la connaissance, parce que je réduis une incertitude dans mes représentations internes. Finalement, il peut y avoir aussi des récompenses de type social, et tu le dis d'ailleurs dans ton article, tu parles de l'empathie, du fait de voir quelqu'un d'autre sourire, et autres. Donc il me semble qu'il y a quelque chose d'assez prometteur dans les directions à creuser dans la façon d'articuler différents objectifs. Comment tu vois les choses là-dessus ?

Yann LeCun [16.40]

Alors, je pense qu'il faut faire une distinction à peu près claire. Les gens qui travaillent sur l'apprentissage par renforcement et qui essaient de l'appliquer au monde réel, à la robotique par exemple, eux voient tout ce que je raconte comme une branche particulière de l'apprentissage par renforcement. Mais moi je trouve que c'est conceptuellement différent. C'est pour ça que je ne l'appelle pas comme ça. Pour moi il y a deux composantes dans l'apprentissage par renforcement. Il y a trois composantes en fait. (1) La première composante, c'est le fait que l'objectif qu'on veut optimiser n'est pas connu. On ne peut pas être rétropropager des gradients à travers. Et les méthodes qui marchent le mieux en apprentissage par renforcement consistent à entraîner un Critic[14], qui est une approximation différentiable de la fonction de coût qu'on veut minimiser, et qu'on entraîne dynamiquement. Donc ça, c'est la première chose. (2) La deuxième, c'est le caractère séquentiel, c'est-à-dire le fait que quand on effectue une action, ça va déterminer ce qu'on va voir [ensuite pour la décision de notre prochaine action]. Donc ce n'est pas comme une situation classique de *machine learning* où les exemples qu'on voit sont indépendants les uns des autres. Là, il y a une dépendance qui est due à ce caractère séquentiel. (3) Et puis la troisième composante, qui n'est pas nécessaire à l'apprentissage par renforcement mais qui est quasi universellement utilisée, c'est l'apprentissage d'une politique, *policy*[15] en anglais : il s'agit de l'apprentissage d'une fonction directe qui à partir de l'état propose une action, ou une distribution sur les actions. Alors là, il y a deux composantes : cette composante de la *policy* est différente de l'idée de planification par MPC (*Model Predictive Control)* par exemple. MPC n'a pas besoin de *policy*. On a juste besoin d'un modèle et d'une méthode d'optimisation qui nous permet de trouver la séquence d'action qui minimise le coût. Si le coût est connu, c'est facile à faire, par méthode à base de gradients. Si le coût n'est pas connu, ou s'il est discret, c'est-à-dire qu'il n'est pas différenciable facilement, on est obligé d'utiliser la recherche arborescente ou une CTS,

14. Dans le domaine de l'apprentissage par renforcement, un modèle *Actor-Critic*, Acteur-Critique en français, contient deux zones mémoires distinctes qui apprennent en tandem : l'Acteur apprend à choisir les bonnes actions au bon moment, c'est-à-dire dans chacun des états de la tâche ; le Critique apprend à prédire la récompense future dans chacun de ses états, et les signaux d'erreur de prédiction de la récompense sont utilisés comme signal de renforcement pour (1) mettre à jour les prédictions du Critique, et (2) mettre à jour les probabilités de choix d'actions de l'Acteur. Le principe se résume ainsi : quand une action donne lieu à quelque chose de mieux qu'attendu (erreur de prédiction positive), cela génère un signal de renforcement positif ; quand une action donne lieu à quelque chose de moins bien que prévu (erreur de prédiction négative), cela génère un signal négatif ; quand une action donne lieu à quelque chose d'attendu, c'est-à-dire bien prévu par le modèle (erreur de prédiction nulle), cela veut dire qu'il n'y a plus besoin d'apprendre et un signal de renforcement nul (sans effet) est généré.
15. Dans l'apprentissage automatique, la politique, *policy* en anglais, est la fonction qui exprime les actions que l'agent a appris à préférer.

de la programmation dynamique, ou l'optimisation combinatoire. Donc ça devient plus compliqué. Enfin là, on retourne dans l'IA classique de recherche de solutions et de planification. Mais si on peut rendre tous les modules différenciables, à ce moment-là on peut utiliser des méthodes de planification par des descentes de gradients. Et là on n'a pas besoin de *policy*. Donc il y a une différence fondamentale entre la commande optimale, *MPC*, et le *reinforcement learning*, *i.e.,* l'apprentissage par renforcement dans lequel on apprend une *policy*.

Je propose de ne pas apprendre de *policy,* sauf… En tout cas pas au début. Bien sûr il y a un phénomène qui se passe chez les humains, et chez les animaux aussi, c'est que quand on fait face à une tâche nouvelle on utilise un modèle du monde, etc. On prête attention à la tâche, on dévoue toute notre attention à cette tâche. On ne peut pas faire autre chose en même temps – par exemple apprendre à conduire, ou à piloter un avion ou même à faire du vélo, du ski. Et puis au bout d'un moment on devient habitué et on peut faire cette tâche de manière subconsciente ; on peut après une cinquantaine d'heures de pratique conduire une voiture et parler à quelqu'un en même temps. Ce n'est pas un problème, alors qu'on ne pouvait pas le faire au début. Ceci tend à suggérer qu'il y a un processus un peu automatique dans le cerveau qui compile une capacité système 2 en une *policy* réactive de type système 1[16]. Et ça, c'est une bonne manière d'engranger des compétences et de pouvoir réagir rapidement et simplement sans avoir à chaque fois à mettre en œuvre notre modèle du monde, qui est un peu compliqué.

Mehdi Khamassi [20.36]

Absolument. Et en neurosciences justement, il y a toute une

branche de recherches qui s'intéresse à comment ça se passe dans le cerveau, comment s'opère la bascule entre ces deux systèmes…

Yann LeCun [20.43]

Oui.

Mehdi Khamassi [20.44]

… dans quelles conditions on considère que ce qu'on a appris, et qu'on a automatisé, est satisfaisant, donc on peut basculer, et quand est-ce qu'il faut au contraire l'abandonner. Et on voit des situations comme ça chez l'humain d'ailleurs. Là on anthropomorphise un peu la discussion en lien à l'IA. Mais vu qu'il y a des liens avec les neurosciences que tu as soulignés, ça peut être intéressant. Il y a en tout cas des fois où nous [humains] persistons dans un automatisme qui n'est plus approprié.

Yann LeCun [21.01]

Oui.

16. La distinction entre système 1 et système 2 a été proposée par le psychologue et Prix Nobel Daniel Kahneman dans son ouvrage « Thinking Fast and Slow » (2011 ; Editions Macmillan) pour schématiser la tendance des humains à alterner entre (au moins) deux modes distincts de prise de décision : un mode intuitif et rapide, où les décisions semblent reposées sur des approximations et heuristiques, qui peut être source d'erreur lorsque l'on répond trop rapidement sans bien réfléchir au problème ; un mode délibératif, où l'on s'approche davantage d'un raisonnement que l'on pourrait qualifier de « rationnel », mais qui nécessite davantage de réflexion que le système 1 et est donc plus lent.

Mehdi Khamassi [21.02]

… Donc il y a des choses intéressantes dont on peut s'inspirer ici.

Yann LeCun [21.04]

Oui.

Troisième partie de l'audition : Hallucinations et autres risques sociétaux

Mehdi Khamassi [0.08]

Bien, alors on a encore un peu de temps. Avant qu'on revienne sur les risques, et tu les as évoqué brièvement, mais quand même aussi, toujours sur le contenu scientifique, j'aimerais qu'on revienne un peu sur les méthodes actuelles qui font beaucoup parler d'elle. Il y a quand même encore des choses à dire. Tu as évoqué très brièvement que dans ces grands modèles de langage il y a ces processus d'hallucination. Est-ce que tu penses que ce type de fragilités pourront être éliminées ? Ou est-ce que c'est quelque chose [les large language models] qui finiront par apparaître comme des gadgets d'intérêt limité ? Qu'est-ce que tu en penses ? Est-ce qu'il y a des directions et un espoir d'éliminer ce genre de problème ?

Yann LeCun [0.44]

Oui, je pense qu'il y a un très bon espoir d'éliminer ces problèmes. Mais à mon avis ça va requérir des changements assez majeurs d'architecture. Donc c'est un petit peu ce que je suggère avec cette idée d'objective-driven AI (en français, IA dirigée vers des objectifs), c'est-à-dire des systèmes qui, au lieu de produire un token après l'autre de manière autorégressive, sans réfléchir à l'avance, planifient leur réponse de manière à satisfaire un certain nombre d'objectifs. Ces objectifs pourraient inclure justement le caractère factuel de la réponse, ou le style de la réponse, ou encore « est-ce que la réponse est compréhensible par un enfant de 13 ans à qui je parle ? », « Est-ce qu'elle est toxique pour une certaine culture ? », etc. Donc à mon avis, avoir un système qui, par construction, produit une sortie qui satisfait un certain nombre d'objectifs permettrait de les rendre pilotables, de les rendre factuels, de les rendre sécurisés, d'éliminer la possibilité de *gear breaking* (en français, rupture de l'engrenage), qui est possible à l'heure actuelle, et probablement donc de réduire, sinon d'éliminer, les problèmes d'hallucination, c'est-à-dire ces problèmes de non-factualité.

Maintenant, il y a les utilisations dans lesquelles on veut que ces systèmes ne soient pas factuels. C'est un peu le même problème qu'on se pose dans les réseaux sociaux : est-ce qu'il faut supprimer la désinformation ? D'abord on n'a pas les techniques d'IA aujourd'hui pour détecter la désinformation. Deuxièmement, on ne sait même pas définir la désinformation, parce que c'est un petit peu dans l'œil de l'observateur, d'une certaine manière, ou en tout cas c'est très difficile à faire sans gros problèmes de censure et de limitation de la liberté d'expression. Et puis il y a aussi le fait que la désinformation n'est pas forcément dangereuse, et dans certains cas, tout à fait désirable : on fait de la poésie, on écrit de la fiction, on se raconte des histoires, des blagues, etc. Tout ça, c'est de la fiction, c'est d'une certaine manière une sorte de désinformation, mais qui n'est pas dangereuse. Donc la

question est : comment détecter et éliminer la désinformation dangereuse, qu'elle soit produite par des humains ou par des systèmes automatiques ?

Aspect « boîte noire » des réseaux de neurones

Mehdi Khamassi [3.09]

Une autre faiblesse qui est souvent pointée du doigt, qui est peut-être plus générale et se rapport à la question des réseaux neurones profonds, concerne l'aspect un peu boîte noire qui peut être assez limitant. Daniel, qui m'a transmis ses questions, disait que quand il en avait parlé avec Yoshua Bengio il y a 3 ans, Yoshua Bengio avait l'air d'écarter un peu ce problème en disant « ça n'a rien de particulièrement inquiétant, ça sera résolu en temps et en heure ». Qu'est-ce que tu en penses ?

Yann LeCun [3.36]

J'ai une opinion encore plus extrême. *(rires)* Je ne pense pas que l'explicabilité soit utile. Les humains ne sont pas explicables. Les chiens encore moins. Encore, les humains peuvent essayer d'expliquer pourquoi, de rationaliser pourquoi ils prennent une décision particulière, ou pourquoi ils ont choisi une séquence d'actions. Pas les chiens. Mais on utilise les chiens pour tout un tas de choses. Il y a tout un tas de choses qu'on fait dans le monde moderne pour lesquelles on n'a pas d'explication particulière, et pourtant on sait que ça marche. Par exemple, on utilise du lithium pour traiter la cyclothymie ou le syndrome à maniacodépressif, on n'a absolument aucune idée de comment ça marche, mais ça marche. Donc l'explicabilité ne sert pas à grand-chose. Elle rassure, mais elle ne sert pas à grand-chose.

Par exemple, je m'inscris en faux contre une loi qui a été passée par la Communauté européenne il y a quelques années sur le fait que toute décision qui est prise par des systèmes automatiques doit être explicable. C'est stupide. C'est d'autant plus stupide que cette loi a dû être modifiée ; on a dû faire un trou dedans pour permettre une autre loi de la Communauté européenne, qui oblige toutes les voitures à avoir un système de détection d'obstacles et de freinage automatique. Ce système s'appelle AEBS : Automatic Emergency Breaking Systems (en français : Système d'Automatique de Freinage en cas d'Urgence). Ce sont des systèmes qui sauvent des vies : ça réduit les collisions de 40% à peu près. Il y a ça dans toutes les voitures qui sortent en Europe maintenant. C'est un réseau convolutif. Le marché est dominé par Mobile AI, qui a 80% du marché. C'est un réseau convolutif, pas explicable, mais ça marche, ça sauve des vies.

Donc il faut se méfier de ce qu'on veut. Votre chauffeur de taxi n'est pas explicable dans la manière dont il conduit. Même un médecin qui fait le diagnostic d'une appendicite, par exemple, en vous appuyant sur le ventre et en essayant de jauger la résistance des muscles dans l'abdomen, n'a pas d'explication à ça. C'est l'expérience qui permet au médecin de faire ça. Ce n'est d'ailleurs pas très fiable. Mais ce n'est pas vraiment explicable.

Si toutes les décisions qui étaient faites par les humains étaient explicables, on pourrait apprendre toutes les compétences en lisant des bouquins. Mais non, ce n'est pas possible. Il faut l'expérience avec le monde réel. Donc ça veut dire qu'il y a plein de choses qu'on fait qui ne sont pas explicables. Et c'est pareil pour les machines ; il faut accepter le fait qu'elles vont être complexes, que les décisions qu'elles vont prendre vont être complexes, et qu'elles sont vraiment difficiles à expliquer. Ceci dit, ce ne sont pas des boîtes noires ; on peut regarder dedans, on peut tout observer. Mais c'est des comportements collectifs émergents. Donc bien sûr c'est compliqué à comprendre.

Mehdi Khamassi [6.28]

C'est une question super importante que tu soulèves aussi sur l'explicabilité. Et au fond je me demande s'il n'y a pas aussi une distinction à faire en fonction des domaines d'application. D'ailleurs c'est quelque chose que tu proposes sur la régulation en général : il faut vraiment faire attention à ne pas faire de la même façon sur différents domaines. Comme tu l'as souligné il y a des exemples où il y a une peut-être une efficacité à sauver des vies, et c'est la priorité, c'est déjà ça.

Parallèle avec l'explicabilité chez l'humain

Mehdi Khamassi [6.49]

Mais j'ai envie de faire un parallèle chez l'humain parce qu'on s'intéresse aussi à la cognition chez l'humain. Et d'une part, quand on est dans une dimension dialogue avec un humain, ce dernier aura tendance à avoir des attentes vis-à-vis de l'agent conversationnel, à lui attribuer des intentions qu'il pourra plaquer dessus, et qui sont les mêmes types d'intentions supposées que peuvent avoir d'autres humains. Donc il faut se méfier de ça. Et d'autre part, il y a cette vision que chez l'humain, l'explicabilité, même si elle n'est pas parfaite – comme le fait d'essayer d'expliquer pourquoi on a pris une décision –, est une dimension, un ingrédient, très important de la cognition sociale.

Yann LeCun [7.20]

Bien sûr.

Mehdi Khamassi [7.21]

Il y a même Dan Sperber et Hugo Mercier qui ont une « théorie argumentative », comme quoi une partie de nos facultés de raisonnement s'est développée principalement pour se justifier dans un contexte social.

Yann LeCun [7.31]

Oui, absolument.

Mehdi Khamassi [7.32]

Donc ça veut dire qu'il y a peut-être des contextes dans lesquels c'est important de pouvoir expliquer pourquoi on a pris telle décision ou telle autre.

Yann LeCun [7.38]

Pour les humains c'est clair que c'est très important, pour la dissémination du savoir, et puis aussi pour établir des relations de confiance. Pour établir que ce que nous dit quelqu'un est fiable, il faut soit (1) que cette personne nous fournisse une explication du raisonnement qui met un peu de lumière

sur la raison pour laquelle la personne nous dit ce qu'elle nous dit – i.e., ce n'est pas simplement ce que l'intuition de la personne suggère, il y a une bonne raison pour tout ça –, ou il faut (2) que ce que la personne en question dit ait été testé encore et encore de manière fiable.

Par exemple, on sait comment volent les avions, plus ou moins. On ne peut pas prédire tous les détails de comment vole un avion ; il y a tout un tas d'explications très compliquées sur ce qui produit la portance d'une aile, et c'est en fait extrêmement compliqué à expliquer à quelqu'un qui n'est pas vraiment féru de physique. On fait tout ça sans prédire la position et la vitesse de chaque molécule d'air. Il n'y en a pas besoin. On peut expliquer qu'un avion ne tombera pas spontanément, qu'il y a tout un tas d'études sur la résistance des matériaux qui font qu'il ne va pas perdre une aile en plein vol, que les turbos réacteurs sont raffinés de telle manière qu'ils ne vont pas exploser ou prendre feu, etc. Ceci fait que maintenant on peut traverser l'Atlantique sur un biréacteur en toute sécurité, de manière complètement fiable. Et donc, on peut essayer de donner des explications de pourquoi c'est fiable – les systèmes redondants, etc., etc., on a des années d'expérience – mais à la fin, la seule chose qui compte, c'est de regarder les statistiques et de voir qu'effectivement il y a un crash pour je ne sais pas combien de millions de milliards de kilomètres parcourus par passager. Et donc c'est incroyablement fiable, beaucoup plus fiable que la voiture, par exemple ! Etc.

Donc il y a deux méthodes là : (1) essayer d'expliquer pourquoi voler dans un avion c'est fiable, ou (2) simplement regarder les statistiques. Et pour une personne qui a peur en avion, c'est toujours regarder les statistiques qui est le plus convaincant. Expliquer pourquoi un avion vole est incroyablement compliqué et requiert des connaissances en génie mécanique, en …

Mehdi Khamassi [10.23]

… en thermodynamique …

Yann LeCun [10.24]

En thermodynamique, etc., et qui sont incroyables ! On se demande comment c'est possible qu'un avion de 400 tonnes puisse tenir en l'air, et comment un turboréacteur peut résister aux chaleurs incroyables qui se trouvent dedans et être si fiable. Et chaque explication qu'on va donner, en fait, va mettre plus de doute dans l'esprit de la personne sur la fiabilité du système. C'est tellement compliqué ; il y a bien quelque chose qui va flancher à un moment ou un autre.

Donc il y a deux manières d'établir la confiance : l'une, c'est l'expérience, et puis l'autre, c'est l'explication. L'expérience est beaucoup plus fiable. Pareil pour les médicaments : on fait des essais cliniques. Expliquer pourquoi les médicaments fonctionnent n'est pas aussi efficace que conduire un test clinique. Jusqu'aux années 70 on ne savait pas comment l'aspirine fonctionnait, et pourtant c'était le médicament plus utilisé au monde.

Mehdi Khamassi [11.19]

Après les choses peuvent se faire dans un deuxième temps, et c'est l'investigation scientifique qui cherche à comprendre pourquoi.

Yann LeCun [11.24]

Bien sûr, et oui c'est toujours nécessaire.

Mehdi Khamassi [11.25]

Et ça peut être absurde de se placer à une mauvaise échelle comme celle des molécules pour comprendre le vol de l'avion. On comprend le principe général, mais à d'autres échelles il y a quand même des explications à trouver et qui peut-être permettront ensuite de passer des idées aux générations futures, technologiquement parlant.

Risques liés à l'IA

Mehdi Khamassi [11.39]

Alors je voudrais qu'on parle des risques. C'est quand même important.

Yann LeCun [11.50]

Oui.

Mehdi Khamassi [11.51]

Effectivement, le débat actuel est sur les risques existentiels potentiellement à 10 ans, 20 ans, 30 ans. Déjà, dans certaines auditions qu'on a réalisées, on en a discuté avec d'autres personnes, et il y a des fois une dimension non fondée, ou parfois un peu romantique. Parfois même il y a une dimension « communication », en fonction de qui annonce ça. Mais ça ne veut pas dire qu'il ne faut pas qu'il y ait un débat, et notamment sur les risques actuels. Il y a déjà des choses actuellement qui posent problème, des choses à court terme. Pour toi, quels sont les risques importants en lien avec l'IA qu'il faut vraiment discuter ?

Yann LeCun [12.13]

Alors, il y a des risques dus aux technologies de communication en général, qui n'ont pas grand-chose à voir avec l'IA, qui existent déjà, mais que les gens associent à l'IA, pour une raison qui est un petit peu mystérieuse. Par exemple, les gens associent des phénomènes qui se sont passés sur les réseaux sociaux dans le passé, et qui se passent beaucoup moins maintenant, à l'utilisation de l'IA, ou en tout cas du *machine learning*. Par exemple, pour les algorithmes de ranking (ordonnancement, en français) sur Facebook, sur Instagram ou sur Youtube, et puis de ranking aussi pour les moteurs de recherche, on a dit « oui, ça conduit à des biais d'information dans lesquels les gens ne voient que les informations qui confirment leurs opinions existantes ». Alors ça, c'était peut-être vrai en 2015, ce n'est plus vrai du tout. Il y a des systèmes de ranking, maintenant, qui sont fait justement pour exposer les gens à des systèmes divers [donc pour favoriser la sérendipité]. Donc ça n'a pas à voir avec l'utilisation du machine learning, d'ailleurs en l'occurrence relativement simple. Ce n'est pas du deep learning (apprentissage profond, en français) de haut vol, parce que ça doit tourner très rapidement des milliards de fois par jour. Donc c'est forcément des choses relativement simples. C'est dû en fait à une bonne définition d'objectifs.

Mehdi Khamassi [13.37]

Oui.

Yann LeCun [13.38]

On en vient toujours un petit peu à ces questions : quel objectif optimiser pour rendre le système utile et pas dangereux à long terme ? Donc il y a déjà ça.

Désinformation

Yann LeCun [13.50]

Ensuite, il y a des gens qui disent que la disponibilité des systèmes tels que les LLMs va permettre à des gens mal intentionnés de disséminer de la désinformation en quantité énorme. Et ça va faire empirer le problème qui existe déjà : en l'occurrence que certaines parties de la population auraient du mal à faire la différence entre l'information et la désinformation. Alors ça, d'abord, il faut répondre que ça existe déjà sans l'IA. C'est-à-dire, par exemple, Qanon, c'est-à-dire des gens qui ont été assez influents aux États-Unis sur les théories complotistes, c'est deux gars. Il n'y a pas d'IA ; c'est juste deux mecs. Et ce qu'ils ont, par contre, c'est un réseau de redistribution de leur contenu. La plupart des contenus, quand ils sont dangereux au niveau des informations, dangereux pour la santé publique par exemple, sont supprimés automatiquement par les réseaux sociaux. Donc ces contenus ne sont pas disséminés énormément. Mais les gens qui veulent trouver l'information vont la trouver, c'est sûr.

Alors, qu'est-ce qui fait que les gens sont sensibles à la désinformation ? On a l'impression qu'ils ont plus sensibles maintenant à la désinformation qu'il ne l'était dans le passé. Il y a pourtant des études très sérieuses qui montrent que ce n'est pas le cas ; les gens ne sont pas plus sensibles à la désinformation maintenant qu'ils ne l'étaient dans le passé. C'est plutôt constant. Mais peut-être que c'est plus visible. Et puis peut-être que dans certains cas les conséquences de ça ont été plus visibles, et donc que les gens se sont plus intéressés à la question. Mais ce sont des études sérieuses, des études académiques de sciences sociales, qui mesurent la sensibilité à la désinformation. Ces études montrent que d'une part c'est essentiellement les gens plutôt âgés qui sont sensibles à la désinformation, pas trop les gens jeunes qui ont grandi avec l'Internet et qui sont beaucoup plus sceptiques, et puis d'autre part, ça n'a pas particulièrement empiré ces dernières années. Donc il faut garder un petit peu la tête froide.

Et puis la deuxième chose surtout, c'est que la modération de contenu sur les réseaux sociaux, par exemple, utilise massivement l'IA. En fait, ça a fait des progrès absolument énormes ces dernières années. C'est-à-dire qu'avant les grands *Transformers* pré-entraînés de manière auto-supervisée, etc., à la Bert, il y a 5 ans sur Facebook, par exemple, la proportion de discours haineux qui étaient supprimés automatiquement par les systèmes d'IA, avant que qui que ce soit ne les voie, était d'environ 25%. Le reste était signalé par les utilisateurs et puis ensuite supprimé manuellement. L'année dernière, c'était 95%. Et la différence, ce n'est rien d'autre que l'utilisation massive de gros *Transformers* préentraînés, self-supervised, multilingues, etc., – pour que ça marche dans toutes les langues du monde, parce que c'est compliqué quand même de détecter le discours haineux dans toutes les langues du monde –. Et puis bon, il y avait beaucoup d'élections, de choses comme ça, un petit peu de courants sociaux, de conflits ethniques et religieux, etc., dans le monde, donc Facebook a décidé d'être assez strict sur la suppression de discours haineux. Cette année, la proportion qui est supprimée automatiquement est un peu plus basse, parce que les tournevis ont été ajustés, parce que le climat est un petit peu moins sulfureux. Donc c'est toujours le compromis entre liberté d'expression et censure.

Mais là, l'IA fait partie de la solution. Ce n'est pas le problème, c'est la solution.

Donc ensuite, évidemment, dans le futur il y a des gens qui vont utiliser les LLMs ou des choses comme ça pour essayer de produire de la désinformation. Le problème de la désinformation n'est pas la production, mais la dissémination. Et en même temps, les entreprises qui ont les meilleurs systèmes d'IA, pour justement détecter ce genre de désinformation, sont justement les entreprises qui sont dans le domaine des réseaux sociaux, des systèmes de communication, etc. C'est Meta, Google, Microsoft, et autres.

Souligner aussi les potentiels positifs de l'IA

Mehdi Khamassi [18.05]

Il y a un truc très important que tu dis en filigrane, et tu as raison, c'est important de le souligner, c'est qu'au fond dès qu'on parle des risques de l'IA, il ne faut pas oublier qu'il y a aussi tout un potentiel de contributions positives à la société.

Yann LeCun [18.18]

Oui.

Mehdi Khamassi [18.19]

Et il faut le rappeler, l'IA peut nous aider à faire des choses, à améliorer. Donc il ne s'agit pas de stigmatiser. Et puis c'est vrai que dans toute la mystification qu'il y a autour de l'IA, il y a ce risque-là, quelque part.

Yann LeCun [18.30]

Oui.

Mehdi Khamassi [18.31]

Donc il faut effectivement se poser pour voir les avantages et les désavantages.

Yann LeCun [18.34]

Surtout en ce qui concerne les progrès de la science, de la médecine, de la sécurité routière, comme je le mentionnais précédemment.

Mehdi Khamassi [18.40]

Voilà !

Yann LeCun [18.41]

Et aussi les systèmes de détection de tumeurs dans les mammogrammes, dans les mammographies. Donc les progrès dans la vie de tous les jours, et puis la préservation de la vie, si on veut, et

puis le bien-être, grâce à l'IA, vont être énormes. Comme pour toute technologie, il y a des risques. Comme utiliser des voitures : les voitures tuent des milliers de personnes par an en France. Mais on est prêt à faire ce compromis. Donc la question est : quels compromis la société est-elle prête à accepter pour l'IA ?

Quels sont les vrais dangers ?

Yann LeCun [19.16]

Et quels sont les vrais dangers ? Beaucoup des dangers de I'IA qui ont été formulés sont des dangers imaginaires. Un en particulier : c'est le danger d'extinction. C'est le scénario à la Terminator dans lequel un beau jour on découvre les secrets de l'intelligence de niveau humain ou surhumain, et puis on allume le système, et puis dans la minute il devient plus intelligent que l'humanité, et il domine l'humanité. Il y a tout un tas de scénarios de science-fiction dans les films ou les romans qui parlent de ça. En fait, ça, c'est complètement improbable.

Mehdi Khamassi [19.56]

Ça peut détourner des risques immédiats et de vraies questions.

Yann LeCun [20.00]

Absolument, absolument. Mais en plus, penser que ces scénarios sont réalistes est une négation de tout ce qu'on connaît sur la manière dont fonctionnent le monde, la technologie, le progrès, etc., etc. C'est-à-dire que c'est un scénario à la James Bond, peut-être.

Mehdi Khamassi [20.19]

Après, voilà, il y aura peut-être des gens qui auront envie de faire ce genre de choses, donc il faut rester vigilant. Mais ce n'est peut-être pas la question la plus importante aujourd'hui.

Possible amplification par l'IA de risques préexistants ?

Mehdi Khamassi [20.26]

Je trouve qu'il y a une deuxième dimension dans ce que tu dis, qui est qu'il y a un certain nombre de risques qui ne sont pas spécifiques à l'IA, qui précèdent l'IA, mais dont on peut quand même se demander, comme pour toute nouvelle technologie qui arrive, s'il n'y a pas des fois une amplification d'un risque existant qui peut être faite par la technologie ? Le fait qu'il y ait de plus en plus de contenu pour lequel ce n'est pas clair si c'est vrai ou c'est faux, ça peut être une des choses amplifiées par l'IA. Le fait que quand on a un processus mono-objectif – par exemple, l'objectif de capter [l'attention de] l'utilisateur, de l'engager –, si on utilise de l'IA pour le profiler, lui proposer des choses qui correspondent bien, on va maximiser encore plus [cette captation de l'attention de l'utilisateur].

Yann LeCun [21.00]

Oui.

Mehdi Khamassi [21.00]

Finalement le problème, c'est le mono-objectif, comme tu l'as souligné.

Yann LeCun [21.02]

Oui. Absolument !

Mehdi Khamassi [21.03]

On peut se dire aussi qu'il y a peut-être une question importante dans le fait d'exploiter des humains pour labelliser des images ou des contenus, et des fois des contenus qui peuvent être très durs à voir, des contenus choquants, étant donné que les algorithmes d'IA ont besoin d'être nourris avec ce type de labellisations. Est-ce que ça n'est pas aussi un problème ? Comment on peut le gérer, selon toi ?

Yann LeCun [21.25]

Je ne sais pas. C'est moins un problème que d'envoyer des gens piocher dans les mines, ou pendant des mois en mer pour conduire un bateau. Enfin, ce sont des métiers difficiles, qui ne sont pas pour tout le monde, mais qui sont proportionnellement plutôt bien payés, en fait, dans les pays où ils se passent. Mais effectivement, l'étiquetage de données, bon, on en a fait plus de cas, je pense, que la magnitude du problème. Il y a effectivement des gens en Inde, en Afrique, qui vivent du fait d'étiqueter des images, des textes, etc., et qui en vivent bien. La raison pour laquelle c'est en Inde, en Afrique, bien sûr, c'est parce que la main-d'œuvre est moins chère, et puis les gens parlent anglais. Donc ça le rend possible. Mais il ne faut pas se voiler la face sur le capitalisme. *(rires)* Mais encore une fois, bon, ce sont des métiers qui… Je préférerais faire ça que creuser du charbon dans une mine. Ça, il n'y a pas de doute ! Ou des diamants en Afrique du Sud ! Enfin, il n'y en a plus beaucoup en Afrique du Sud. Donc, ce n'est pas pour tout le monde ; il faut effectivement avoir les nerfs assez solides.

Mehdi Khamassi [22.47]

Et sans que ça soit un problème spécifique à l'IA, peut-être même que le fait que ça soit remis en lumière avec l'utilisation de l'IA est une occasion renouveler de discuter de cette question d'exploitation, ou de certains problèmes comme ça.

Yann LeCun [23.00]

De l'organisation économique du monde, de la disparité entre les pays !

Mehdi Khamassi [23.04]

Voilà ! Oui, il faut discuter de ces choses.

Yann LeCun [23.06]

Ce sont des problèmes qui ne vont ni être magnifiés ni être résolus par l'IA, j'en ai peur. Ce sont des problèmes politiques.

Mehdi Khamassi [23.14]

Peut-être que l'IA pourrait nous aider à trouver des solutions alternatives.

Yann LeCun [23.16]

C'est possible.

Questions environnementales et coût énergétique de l'IA

Mehdi Khamassi [23.19]

Il y a un autre risque qui me paraît important, avec les questions liées à l'environnement, le réchauffement climatique, les coûts énergétiques. Au fond, c'est vrai que nous tous, chercheurs en IA, en robotique, utilisant des calculs intensifs, on doit se poser la question de l'empreinte énergétique. Il y a certains de ces algorithmes, notamment des grands modèles, qui ont besoin d'énormément d'énergie. Donc, comment tu vois les choses ? Qu'est-ce qui peut être amélioré de ce côté-là ?

Yann LeCun [23.46]

Alors, déjà, disons, dans l'image complète de la consommation énergétique dans le monde, tout ça c'est relativement faible, c'est quasiment négligeable. C'est la première chose. La deuxième chose qu'il faut savoir est qu'il y a une incitation absolument gigantesque pour les grandes entreprises à minimiser l'énergie qui est dépensée dans ces systèmes. Parce que c'est un poste principal de leurs dépenses. Et la conséquence de ça, qui est intéressante, c'est que la proportion de l'énergie électrique consommée dans les technologies de l'information, d'une manière générale, et dans les centres de calculs, les centres de données et dans l'IA, en particulier, est constante ; elle n'augmente pas. C'est-à-dire que la totalité de toute la technologie de l'information en consommation électrique, c'est à peu près 6% de l'énergie totale dans le monde. Là-dedans, les datacenters, qui sont utilisés par Google, Facebook, et toute l'industrie, c'est 2 à 3% de l'énergie totale. Donc, ce n'est pas négligeable, mais ce n'est pas gigantesque, et c'est constant, ça n'augmente pas. La quantité de calcul augmente exponentiellement, parce que l'efficacité des calculs augmente exponentiellement, par unité d'énergie dépensée. Mais la quantité d'énergie consommée là-dedans est constante.

Et la raison pour laquelle elle est constante est économique. C'est-à-dire que si vous êtes français, que vous utilisez Facebook ou Instagram, Meta va gagner peut-être 10 à 15 euros par an avec les pubs sur lesquelles vous avez cliqué une fois de temps en temps. Donc Meta ne peut pas dépenser plus que quelques euros par an en énergie pour les services rendus. C'est limité par ça. Et la proportion de revenus par Meta par personne, n'augmente pas particulièrement. Le nombre de personnes augmente un peu, mais pas très vite, plus très vite. D'ailleurs, elle a même décru en 2022. Mais les revenus n'augmentent pas à une vitesse énorme. Et donc cette énergie consommée est constante. Ça, c'est la première chose.

La deuxième chose est que les opérations de Facebook sont neutres au niveau carbone. Facebook est le l'acheteur le plus important de l'énergie renouvelable aux États-Unis. Bien sûr, là-dedans il y a des trades (des échanges, en français), parce qu'on a besoin d'énergie solaire là où il n'y a pas de soleil, etc. Voilà.

Utiliser l'IA pour trouver des solutions au changement climatique

Yann LeCun [26.37]

Et puis, il y a des projets qui utilisent l'IA, maintenant, dont l'objectif principal est d'essayer de trouver des solutions au changement climatique. Par exemple, un projet qui a été monté par mes collègues à FAIR [Facebook Artificial Intelligence Research, centre de recherches à Paris], qui s'appelle opencatalystproject.org. Et ce projet consiste à utiliser les supers ordinateurs de Meta pour faire des calculs de dynamique moléculaire, pour essayer de prédire quelle est l'interaction entre une molécule d'eau et un substrat, un catalyseur, etc., avec des composés connus, et utiliser ces bases de données pour entraîner un système d'apprentissage, de Deep learning, d'IA, pour essayer de trouver de nouveaux composés qui permettraient de catalyser la réaction de séparation d'hydrogène et de l'oxygène avec l'électricité. Parce que si on pouvait avoir un système comme ça pour faire de l'hydrogène à partir de l'électricité, on pourrait tapisser un petit désert de panneaux solaires, et puis stocker l'énergie et l'envoyer là où on en a besoin sous forme d'hydrogène ou de méthane. Ce serait un pas énorme vers la résolution du changement climatique.

On ne sait pas si ça va marcher. C'est un projet de recherche. Recherche ouverte et collaborative, d'ailleurs. N'importe qui peut participer, les données sont disponibles de manière ouverte et gratuite. Il y a une espèce de *leader board* [*i.e.,* un tableau comparateur de performance], donc de compétition constante dans laquelle les gens peuvent participer pour essayer de trouver les meilleurs catalyseurs.

Et c'est important d'essayer de résoudre cette question parce qu'on sait faire la séparation de l'hydrogène et de l'oxygène, mais les méthodes qu'on a sont soit efficaces, soit peu chères et il est donc possible de passer à l'échelle, mais pas les deux à la fois. Les catalyseurs efficaces, c'est du platine, c'est beaucoup trop cher, on ne peut pas passer à l'échelle. Et puis sans catalyseur, ce n'est pas efficace. Donc voilà.

Mehdi Khamassi [28.46]

Nul doute que les techniques d'IA ont un potentiel énorme de nous aider à résoudre énormément de problèmes, à mieux comprendre le monde.

Affaiblissement des universités face aux géants du Web

Mehdi Khamassi [28.54]

Alors, j'ai une dernière question à te poser. Je crois qu'on a encore quelques minutes, à peine. Quelque chose qui pourrait être vu un peu comme un risque, mais c'est aussi une question sur l'or-

ganisation de la recherche en IA. Et ce n'est pas spécifique à l'IA, ça s'est passé avant avec d'autres technologies, mais en tout cas, en ce moment, beaucoup de choses et des moyens pour avancer sur la recherche en IA sont beaucoup plus forts chez les grands industriels, les géants du Web. Et il y a un affaiblissement des universités. Alors, il y a un certain nombre de collègues qui, comme toi finalement, gardent un pied dans une université et l'autre pied dans le milieu industriel. Quelle complémentarité cela permet de faire ? Et qu'est-ce que tu crois qu'il faudrait faire pour les universités dans ce contexte ?

Yann LeCun [29.27]

Alors oui, la raison pour laquelle j'ai quand même gardé un pied à l'université est que je pense que les recherches qui sont faites en industrie et en université sont complémentaires. Il y a beaucoup de bonnes idées qui viennent des universités et qui n'ont pas nécessairement besoin d'énormes moyens de calcul pour être montrées, si ce n'est que sur des problèmes jouets. Un bon exemple de ça ces dernières années est toute l'idée d'attention dans les réseaux de neurones, qui vient du laboratoire du Yoshua Bengio à Montréal, avec un stagiaire en Master et un post-doc, à l'époque, qui s'appelle Kyunghyun Cho. C'est maintenant un collègue à NYU. Et cette idée s'est répandue comme une traînée de poudre, et a permis de faire des systèmes de traduction, par exemple, aux alentours de 2015-2016, qui marchaient bien mieux que ce qu'on avait avant. La raison pour laquelle vous voyez de bonnes traductions maintenant est en partie due à cet impact. Et puis ça a donné naissance aux architectures *Transformers*, l'idée étant que si on se repose sur ces méthodes d'attention de manière assez massive, on obtient des réseaux avec des propriétés intéressantes.

Donc les progrès qu'on a vus récemment viennent d'une recherche à l'origine universitaire. Beaucoup de bonnes idées viennent des universités, comme par exemple de nouvelles manières de faire les choses. L'industrie tend à être beaucoup plus pilotée par des modes, parce que quand on développe des produits il faut aller avec ce qui marche. Donc on a un petit peu moins de temps pour lever le nez du guidon.

Ce qu'on a un Meta, c'est une espèce de fusée à plusieurs étages : il y a le FAIR qui fait de la recherche fondamentale. Le FAIR est lui-même divisé en deux sous-labo (un qui s'appelle FAIR Labs, qui est vraiment de la recherche bottom-up, déterminée par les chercheurs eux-mêmes, etc., avec des petites équipes, pas trop d'ingénieurs ; et puis FAIR XL qui travaille avec des projets un peu plus organisés, un peu plus gros, avec plus de supports, d'ingénieurs, plus de moyens de calcul). Et puis ensuite on a des groupes de développement avancé de produits, qui font passer l'échelle, qui font marcher les choses, etc., et qui ont beaucoup moins de temps de faire de la recherche et d'essayer de changer les choses. Donc il y a besoin un peu de tout ça. Et puis, en collaboration avec FAIR Labs, beaucoup de collaborations avec les universités, comme par exemple à Paris on a une trentaine d'étudiants en doctorat en résidence CIFRE[17], et qui font un travail extraordinaire et qui essaime l'écosystème français de R&D en IA d'ailleurs. On est peut-être la plus grosse école doctorale d'IA en France, bizarrement. *(rires)* Et tout ça bien sûr est en collaboration avec des labos universitaires, avec l'INRIA, etc., qui sont co-auteurs de tous les papiers.

Mais bon, ça pose un vrai problème qui est que si on s'intéresse, dans une université, à une application telle que, je ne sais pas, des applications en vision, ou des applications bien connues en reconnaissance de la parole, en traduction, en systèmes de dialogue, on ne va pas battre les records des systèmes industriels dans lesquels il y a des équipes énormes, des moyens de calcul monstrueux,

17. Les Conventions Industrielles de Formation par la REcherche (CIFRE) sont un dispositif qui existe depuis de nombreuses années en France pour permettre à des doctorants d'effectuer une thèse impliquant une collaboration entre une entreprise privée et une université ou un organisme public de recherche.

etc. Il ne faut pas se mettre sur ce terrain. Il faut produire de nouvelles idées, de nouvelles manières de faire les choses, essayer d'imaginer le futur, faire un peu de théorie parce qu'on en manque. Mais il faut quand même que les gouvernements mettent à disposition des universitaires des moyens calculs non négligeables. En France, on a eu un peu la chance d'avoir accès au calculateur Jean Zay, qui était un service assez tôt et qui a permis à la recherche universitaire française d'utiliser des GPU relativement rapidement. Et puis là, il y a une deuxième vague de systèmes de calcul qui va être faite. Les États-Unis sont un peu en retard. Ils commencent à réaliser qu'il faudrait mettre des moyens à disposition des universitaires, et d'autres pays aussi [commencent à le réaliser]. Je pense que ça, c'est la chose essentielle.

Mehdi Khamassi [33.40]

Vraiment un grand merci, Yann ! C'était super intéressant ! Plein de bonnes choses à toi, passe un bel été et puis à bientôt !

Yann LeCun [33.48]

Merci. À bientôt !

Utilisation de la robotique sociale et du jeu sérieux en psychiatrie

Audition de David Cohen

DAVID COHEN

David Cohen est professeur à Sorbonne Université et directeur du service de psychiatrie de l'enfant et de l'adolescent à l'hôpital de la Pitié Salpêtrière (APHP). Il est membre de la nouvelle équipe Action, Cognition, Interaction et Décision Encorporées de l'Institut des systèmes intelligents et robotique (ISIR) à Paris (CNRS / Sorbonne Université). Il est spécialiste des troubles de l'apprentissage, et en particulier de l'autisme, mais aussi de schizophrénie, catatonie et troubles sévères de l'humeur apparaissant pendant l'enfance. Il a été président de International Association of Child and Adolescent Psychiatry and Allied Professions (IACAPAP) 2012 congress.

L'audition a été menée par Mehdi Khamassi et Florian Forestier

Mehdi Khamassi

Bonjour David Cohen. Merci beaucoup d'avoir accepté de répondre à nos questions pour TESaCo. Pour commencer, pourrais-tu te présenter et nous dire quelques mots sur tes responsabilités, et ton parcours ?

David Cohen

Merci Mehdi pour l'invitation. Je suis psychiatre de l'enfant et de l'adolescent à l'hôpital de la Salpêtrière et à Sorbonne Université. J'ai une activité de recherche en partenariat avec l'Institut des systèmes intelligents et robotique (ISIR) dans une équipe qui s'appelle "Perception, Interaction et Robotique sociale". Notre but est à la fois de travailler sur ce que les psychiatres ont du mal à voir ou à interpréter, c'est-à-dire des signaux sociaux relativement subtils qu'on ne voit pas à l'œil nu ; et de développer des applications dans le champ thérapeutique, soit par l'accompagnement en robotique, soit en utilisant ce qu'on appelle des jeux sérieux et éducatifs qui permettent de travailler un certain nombre de problèmes que les enfants pourraient présenter au cours de leur développement.

Je crois qu'en tant que psychiatres, nous apportons aux roboticiens une certaine forme d'« expertise » en ce qui concerne les émotions, les interactions sociales, les interactions en petits groupes, les questions d'affiliation, ou pas, au cours des interactions sociales. Donc une expertise en psychologie sociale. Nous apportons également une expertise en psychiatrie : avoir accès à des pathologies qui peuvent devenir des modèles extrêmes de dimensions pertinentes qu'on pourrait étudier. Pour les interactions sociales, tout le monde comprend que les enfants autistes, du fait de la nature de leurs interactions sociales, sont finalement l'exemplarité des difficultés les plus grandes dans cette dimension.

C'est le cas en particulier quand on veut tester certains modèles. Mais ça peut être vrai aussi pour le langage, puisqu'on a des enfants avec des troubles du développement du langage. Ça peut aussi être vrai pour la motricité, puisqu›on a des enfants avec des troubles de la coordination motrice. Tout ceci va éclairer les algorithmes, qui peuvent être produits en robotique ou en traitement du signal social, de la norme aux des personnes en plus grandes difficultés. Ça permet de mieux stabiliser les modèles.

Utilisation de la robotique sociale et du jeu sérieux en psychiatrie

Mehdi Khamassi

Super ! Ça me permet de faire le lien avec TESaCo, qui coordonne cette audition, et qui est le projet « Technologies émergentes et sagesse collective que Daniel Andler pilote pour l'Académie des sciences morales et politiques. Dans ce projet, nous nous posons des questions sur ces nouvelles technologies émergentes, qui ont l'air d'infuser assez rapidement dans la société, et qui ont l'air de faire un effet de rebond les unes sur les autres pour faciliter les choses. Nous nous demandons : où cela emmène la société ? Qu'est-ce que cela nous permet de développer comme applications ? Quelles peuvent être les risques ? Dans ce cadre, j'aimerais que tu en dises un peu plus sur ton activité qui est spécifiquement en lien avec les technologies émergentes, en particulier les robots, mais pas seulement.

David Cohen

Les activités que nous avons en robotique sont de deux ordres. Il y a d'une part certains modèles qui concernent essentiellement le traitement de signal social, que l'on va tester sur des questions, souvent partagées avec les ingénieurs en robotique sociale. Par exemple, il s'agit de mesurer des signaux extrêmement subtils dans la gestion du stress, dans les variations d'émotion, dans le rythme des interactions entre personnes, donc des choses de cet ordre. Le deuxième type d'activité, ce sont des activités d'applications thérapeutiques. Je dirais que nous avons deux grands types d'applications thérapeutiques : d'une part la robotique d'accompagnement. Par exemple, dans notre protocole actuel, les deux principales activités que nous essayons de développer, et que je vais présenter comme un fantasme même si ce n'est pas encore réellement opérationnel, consistent à travailler avec des écoles ou avec le monde du handicap pour insérer un enfant qui a des difficultés importantes de développement à l'école. Souvent, il y a des aides de vie scolaire, des personnes qui viennent spécifiquement pour aider ces enfants. Nous souhaiterions faire des robots d'accompagnement qui seraient des robots AVS [assistants de vie scolaire], non pas pour remplacer les AVS en eux-mêmes, car ce n'est pas demain que les robots remplaceront les humains, mais au moins que les robots puissent faire des tâches assez spécifiques. Ceci requiert que ces robots soient conçus avec des capacités d'adaptation suffisante, ce qui est un des grands enjeux actuels de la robotique. Avec de telles capacités, les robots pourraient tout à fait accompagner de manière plus performante certaines problématiques, notamment en détectant des indices de compréhension d'interaction qui ne sont pas visibles par l'humain à l›œil nu. En ce moment, nous avons deux projets qui vont dans cette direction-là : un sur l'écriture et un sur l'engagement attentionnel.

Le deuxième type d'applications que nous développons actuellement avec un certain succès, car on avance avec des process de plus en plus performants, est ce qu'on appelle les jeux éducatifs sé-

rieux. Ce sont des jeux un peu comme ceux que n'importe quel enfant peut avoir sur des tablettes, des ordinateurs, des téléphones ou des consoles. Mais nous introduisons dans le jeu une sorte d'agenda éducatif qui existe à l'insu du joueur et qui va l'entraîner, en jouant, à perfectionner une activité ou une dimension qui est en difficulté. Ce qui est intéressant dans les *serious games* les plus performants, c'est qu'on peut même introduire à l›intérieur du programme éducatif des algorithmes qui s'adaptent aux spécificités des joueurs, aux spécificités des dimensions qu'on veut travailler. Ceci donne des jeux qui ont une qualité d'adaptation à la personne qui est absolument extraordinaire. Encore une fois, cela se passe à l'insu de l'enfant. Ce point est important, car les enfants ont besoin de rééducation. Souvent, quand on leur dit « je vais prendre un exemple lié à l'écriture », l'enfant va nous répondre qu'il n'aime pas l'écriture. Nous avons beaucoup travaillé sur ce sujet, et avons rencontré des exemples d'enfants que nous voulions faire progresser en écriture parce qu'ils avaient des problèmes de coordination motrice. C'est alors normal que l'enfant refuse, puisqu'il n'arrive pas très bien à écrire. Les enfants sont assez pragmatiques ; ce qui ne marche pas très bien, ils veulent en général l'éviter. Le problème est que si on leur dit « tu vas aller voir un graphothérapeute ou un psychomotricien pour travailler l'écriture », certains nous disent qu'ils en ont marre. Si par contre on leur dit « tu vas jouer à un jeu, telle ou telle activité, avec un stylet sur une tablette », alors il ne s'agit plus explicitement d'écriture, mais de jeux qui vont travailler un certain aspect de l'écriture sur un mode complètement ludique. Là, il n'y a plus de problème de motivation, ni de problème d'opposition : on peut le diriger vers des exercices de rééducation qui deviennent beaucoup plus confortables, car beaucoup plus ludiques.

Évolution récente du champ vers la psychiatrie computationnelle

Mehdi Khamassi [vers 9 min]

Est-ce que ce type d'utilisations de robots ou de jeux sérieux traduit une évolution générale en psychiatrie ? Ou ce sont des choses spécifiques aux recherches de certains labos et de certains chercheurs ? Comment vois-tu l'évolution actuelle en psychiatrie ?

David Cohen [9:39]

Dans l'évolution actuelle de la psychiatrie, il y a un grand courant assez nouveau dont l'expression anglo-saxonne est « Computational psychiatry », c'est-à-dire la psychiatrie computationnelle. Dans la psychiatrie computationnelle, il y a une partie des choses qui ne correspondent pas forcément à mes activités prioritaires et qui concernent la gestion des *Big data*. Je trouve d'ailleurs qu'il y a une sorte de confusion dans le grand public. Cette confusion est en partie due à l'utilisation du concept d'*intelligence artificielle*, qui est devenu un concept avec beaucoup de polysémie ; les gens utilisent le terme « IA » parfois avec une tendance très expansive. Je pense que c'est lié au succès de la méthode et à sa facilité d'accès aujourd'hui. Alors qu'à une époque il fallait quand même des gens qui étaient un peu experts pour le faire, et surtout c'était un domaine beaucoup moins accessible par le grand public, car ce dernier n'avait pas accès à des machines aussi puissantes qu'aujourd'hui. Depuis, il y a eu une forme de démocratisation. Même une personne qui ne comprend rien peu, avec un petit partenariat, s'y mettre et faire des choses. On voit avec les *Big data*, dans tous les champs de la science, que c'est devenu un sujet en soi. Ceci est vrai aussi en psychiatrie. Donc certains vont analyser d'énormes

bases de données, ou vont essayer de trouver des modèles qui répondent à de nouvelles questions. À partir des analyses de ces bases de données, on peut faire du *machine learning* pour essayer de faire de la prédiction, et ce même si on ne peut pas améliorer un certain nombre de choses, ni faire de la médecine dite individuelle ; on peut néanmoins essayer sur des données géantes ou macro.

Finalement, ça c'est une recherche en psychiatrie qui existe aujourd'hui, et il y a pas mal de gens qui travaillent dans le domaine. Et c'est probablement là qu'il y a le plus d'équipes. De même, dans la robotique sociale et dans le jeu sérieux, il y a pas mal de projets collaboratifs ; c'est très porté par les agences de financement, qui veulent de l'interdisciplinaire – on connaît tous ces mots clefs qu'on nous sert. Mais les projets collaboratifs ne sont pas forcément multidisciplinaires, et ce sont rarement des projets intégrés. Or, de mon point de vue, c'est justement le type de travaux que nous parvenons à mener à l'ISIR – car l'ISIR est un laboratoire de robotique et d'intelligence artificielle qui en son sein héberge un certain nombre de psychologues. Ceux-ci peuvent ainsi travailler de manière très intégrée, avec de véritables allers-retours entre les questions psychologiques et les méthodes plus informatiques et mathématiques.

Imitation enfant-robot

Même si on comprend bien l'évolution des différents modèles, et éventuellement les questions qu'on se pose à l'interface entre psychologie et ingénierie, il y a toujours des surprises. Et ça, c'est extrêmement fécond. Je peux prendre un exemple de cette fécondité d'intégration du côté des travaux que nous avons faits sur l'imitation en robotique. Nous menons en effet des travaux d'imitation enfant-robot pour essayer de comprendre quelque chose du rythme d'interaction[1]. Nous travaillons avec un modèle qui est capable d'apprendre par imitation[2]. Nous pouvons alors étudier comment l'interaction avec de tels modèles fonctionne, et comparer ce qui se passe avec les enfants ou avec des adultes, avec des enfants autistes ou avec des enfants bien portants. Est-ce que le type de population change la manière d'apprendre pour le robot ? Nous avons constaté en effet qu'il y a des différences. Puis en regardant dans le détail ces différences, en particulier ce qui change dans les réseaux de neurones qui sont activés avec différents participants, nous nous sommes rendu compte qu'il y avait des petites particularités avec les autistes qui sont assez questionnantes. En particulier, comme il s'agissait d'une tâche d'imitation motrice, nous avons clairement vu, en analysant les résultats à deux, ingénieurs et psychologues, que l'adaptation passe par la détection par le modèle d'une sorte de signature motrice. Ceci nous permet de perfectionner notre modèle avec un nouveau questionnement autour de la reconnaissance de la signature motrice de l'autre. Nous avons alors refait les expériences et avons vu que le modèle est encore plus performant. Ceci confirme qu'en apprenant par imitation on peut reconnaître la signature de l'autre, alors que ce n'était pas prévu par le modèle (en tout cas tel que l'expérience était faite). Donc c'était un joli résultat. Et maintenant cela nous amène à nous demander comment évoluent ces signatures motrices au cours du développement chez l'enfant autiste (pour des raisons physiologiques que je pourrais expliquer par ailleurs, qui ne sont pas en lien avec la question, mais qui sont importantes parce que si cela valide une question physiologique, c'est que probablement cela a du sens au global). Ceci a ainsi débouché sur toute une série d'expériences où nous allons chercher chez les enfants autistes, au cours du développement, s'il y a une persistance de

1. NdE : L'intérêt du robot est alors d'avoir un comportement parfaitement contrôlé, et de pouvoir étudier comment l'enfant réagit ou s'adapte quand le robot a tel ou tel comportement prédéfini.
2. NdE : Donc ici, le robot n'a pas seulement des comportements prédéfinis, mais des comportements qui s'adaptent aux réactions des différents enfants.

ce que l'on appelle les micromouvements, ou submouvements, qui sont présents au cours du développement pendant un mouvement volontaire. Nous nous demandons, et c'est pour cela que j'ai utilisé le mot *micromouvement*, s'il s'agit simplement d'une subsistance plus longue du mouvement chez l'enfant autiste, ou si c'est la nature même de ces submouvements qui est modifiée, voire déviante. Et à ce moment-là on va les appeler plutôt *micromouvements*, pour les séparer des submouvements, qui sont classiquement décrits au cours de l'apprentissage de mouvements volontaires. C'est clairement quelque chose que l'on n'aurait pas fait dans le cadre d'un rapport de consortium, parce que dans un consortium on collabore sur une recherche, on fait le projet, et puis quand le projet est terminé, on en reste là. Ce n'est pas pareil quand la recherche est intégrée dans une même équipe de recherche. Cela fait 10 ans que je suis à l'ISIR, et j'ai vu cette évolution. Pour revenir à la question « est-ce que c'est très à la mode en psychiatrie ? », je dirais que les équipes intégrées, non, il n'y en a pas beaucoup. Par contre, le fait qu'il y ait tout un courant de nouvelles technologies en psychiatrie, donc ce qui est regroupé sous le terme « computationnal psychiatry », c'est clair que c'est un vrai sujet. Mais je crois que nous n'avons pas encore un journal dédié. En général, quand un sujet en psychiatrie devient à la mode, on en fait un journal.

Mehdi Khamassi

Il me semble que même en psychiatrie computationnelle, sans utilisation de robot, mais comme tu le décrivais, dans un champ où l'on fait des modèles mathématiques, où l'on analyse des données sur des tâches notamment comportementales dans différentes pathologies, il y a aussi un nombre d'équipes en France qui se comptent sur les doigts de la main ; c'est encore balbutiant. Ce que tu décris avec cette dimension intégrée, incluant des ingénieurs, me semble moins une grande tendance de la psychiatrie en termes d'applications aux soins, si je le comprends bien, qu'un domaine de recherche encore très exploratoire.

Neurosciences computationnelles appliquées à la psychiatrie

David Cohen

Oui, c'est sûr. Alors, il y a un autre domaine que je n'ai pas cité, qui viendrait plutôt des neurosciences appliquées que de la psychiatrie : les neurosciences computationnelles appliquées à la psychiatrie. Ce domaine est fondé sur des connaissances de neurosciences computationnelles qui portent en particulier sur des architectures fonctionnelles. L'application en psychiatrie reste relativement limitée. Ça fonctionne très bien pour un certain nombre de pathologies neurologiques dans ce domaine-là. Mais en psychiatrie cela reste quand même balbutiant. Probablement parce que les modèles anatomo-fonctionnels en psychiatrie sont moins robustes que les modèles en neurologie. Mon hypothèse est qu'en psychiatrie il y a une dimension de complexité, même sur le plan phénoménologique, qui n›est pas celle de la neurologie. Donc l'idée d'avoir une corrélation anatomo-clinique comme en neurologie, est, à mon avis, un peu illusoire. Néanmoins, c'est un courant qui est assez présent aussi dans les neurosciences computationnelles, par ce que c'est un enjeu pour les neurosciences computationnelles et qu'il amène aux questions suivantes : est-ce qu'en ajoutant une puissance de calcul supplémentaire on va pouvoir faire un saut conceptuel ? Est-ce que cela va permettre de passer d'une compréhension neuropsychologique ou neurologique, on va dire, dans un certain nombre

de phénomènes ou de maladies neurologiques, à une compréhension plus complexe mais qui serait possiblement applicable à un certain nombre de problématiques psychiatriques ?

Questions éthiques

Mehdi Khamassi

À partir de la description du paysage scientifique que tu viens de faire, je voudrais t'interroger finalement sur les interactions que tu vois entre la dimension recherche, la dimension éducation et la dimension soin. Est-ce qu'il y aurait des spécificités de ces interactions dans le cas de soins pour les mineurs, et notamment pour les mineurs atteints de troubles psychiatriques ?

David Cohen

Par rapport aux questions éthiques de recherche, au stade où on en est, il y a encore beaucoup de travail avant de pouvoir bien les identifier toutes. Il faut bien voir qu'aujourd'hui il s'agit de recherches qui sont extrêmement exploratoires, en particulier en robotique. En robotique, quand on a une expérience avec 10 enfants, c'est énormément de travail, de mise en œuvre. Le plus souvent, quand on voit une vidéo avec un enfant qui interagit avec un robot, ce que les gens ne comprennent pas, c'est que le robot est contrôlé par un opérateur. Donc il ne s'agit pas d'une autonomie réelle. Et puis ce sont des montages, donc on voit les moments les plus sympas. Mais tant s'en faut c'est une interaction continue d'une certaine durée. Aujourd'hui, je ne crois pas qu'il y ait de problème majeur en termes de données, de big data, c'est-à-dire en lien avec les problématiques éthiques spécifiques autres que l'éthique de recherche. Pour les *serious games*, le cas est un peu différent. D'abord parce que les jeux vidéo existent en tant que marché. On peut donc imaginer que les jeux éducatifs sont une application particulière du marché. Donc, les problématiques qui vont se poser sur le plan des soins sont les mêmes en termes de contrôle des données que tout ce qui concerne la présence sur le Web. En effet, la plupart des jeux d'aujourd'hui ne se font pas avec un disque qu'on met dans une plateforme physique, comme au début du jeu vidéo. Ça se fait le plus souvent par un accès en ligne, où on va avoir la prestation qui apparaît à l'écran, etc. Dans ce cas-là se posent les mêmes problèmes de données parce que c'est forcément des serveurs, souvent que des GAFA mettent à disposition. Même quand on a une petite structure, on est obligé de passer par le réseau Internet, qui est un réseau mondialisé, globalisé. Cela n'est pas spécifique à ce à quoi on s'intéresse. Par contre, il y a une question éthique autour du contrôle qualité du jeu. Car aujourd'hui rien n'interdit à quelqu'un qui fait un jeu éducatif de dire : « moi je vends un jeu qui est magnifique pour l'apprentissage de la lecture, donc je dis que c'est bien pour les dyslexiques ». Il n'y a pas d'autorité qui va contrôler si réellement ce jeu a prouvé un intérêt pour les enfants dyslexiques. Avec les jeux sérieux, on se situe un peu comme avec les plantes, pour faire un parallèle avec la psychopharmacologie. Par exemple, demain on va dans une pharmacie, on trouve la plante qui aide à dormir, des pierres qui aident à trouver un certain karma, et c'est en vente libre ; les gens qui ont envie de les acheter les achètent, etc. C'est un vrai problème, car le public et les médecins, ou les psychologues, les éducateurs, les enseignants, ne savent pas forcément quel jeu a montré une qualité spécifique et robuste pour leur objet d'intérêt. Il est clair quand on regarde le nombre de jeux disponibles et le nombre d'études qui ont été faites pour regarder l'intérêt clinique de ces jeux dans les pathologies qui réclament un intérêt, et bien là on se rend compte qu'il y a un fossé énorme. Certaines équipes qui travaillent sur ces jeux l'ont fait, mais c'est loin d'être une généralité.

Actuellement, il y a des régulations. Par exemple, il y a quelques jeux qui ont été enregistrés à la FDA[3] comme étant des jeux qui ont suffisamment d'études cliniques pour penser qu'ils sont pertinents dans telle ou telle problématique, et qui sont donc remboursables par des assurances privées. En pédopsychiatrie, ça a été le cas récemment pour un jeu sur les troubles de l'attention (c'est tout un marché parce qu'il y a beaucoup d'enfants avec des troubles de l'attention et d'hyperactivité). Et je crois que la Commission européenne est en train de revoir toute une série de réglementations pour faire une place de manière plus globale à l'e-santé, parce que ce ne sont pas forcément que des jeux sérieux, que des outils d'e-santé, qui doivent avoir un label de vérification médicale.

Dans le champ de l'autisme, je ne trouve pas de problème éthique de fond, si ce n'est des problèmes de sécurité. En effet, une machine, un robot, ça pèse lourd, et il faut donc faire attention.

Implication des familles d'enfants avec troubles d'apprentissage

Mehdi Khamassi

Avant de pousser sur ces questions éthiques, et de la manière dont cela s'infuse dans la société, je voulais, toujours sur ses histoires d'interaction entre recherche, santé, éducation et soin, dans le cadre de tes travaux, savoir quel est le rôle ou quelle est l'intégration qu'il y a pour entourage familial dans le choix des participants et dans le suivi ?

David Cohen

Déjà, il y a le cas général de la recherche. Toutes les recherches sur l'enfant et l'adolescent se font avec l'accord des parents, puisqu'ils sont mineurs. Les deux parents doivent donner leur accord en général. De plus, dans nos travaux nous sommes parfois extrêmement inclusifs de l'aspect familial. Pour prendre un exemple, nous avons même développé un jeu sérieux dont l'intérêt dans la conception du jeu est l›interaction entre parents et enfants. C›est un jeu qu›on a construit pour les enfants autistes, et qu›on est d›ailleurs en train de tester à grande échelle. On attend juste l›enregistrement européen du jeu pour démarrer l'étude « e-Golia ». Il est mis à disposition sur une plate-forme qui a un enregistrement d'e-santé. Mais avant ça, le jeu doit être enregistré. Pour revenir à l'intégration des parents, ce jeu fait travailler un certain nombre de choses chez l'enfant autiste, comme l'imitation ou l'attention conjointe. Mais la construction du jeu fait que l'enfant ne peut pas jouer seul. Pour y jouer, il est obligé d'interagir avec un proche (le plus souvent un parent), puisque c'est un jeu où il faut deux tablettes connectées pour jouer. C'est un peu comme si on se renvoyait la balle à travers Internet pour pouvoir jouer. Donc, ça, c'est pour moi un élément assez important, puisque ça intègre l'interaction avec les parents.

La première étude pilote que nous avions faite, et toutes les questions que nous nous étions posées sur le plan éthique étaient orientées vers notre idée de promouvoir des actions thérapeutiques même à la maison, quand les enfants sont avec les parents. Mais nous nous étions dit qu'après tout, transférer la charge des soins à la maison pour les parents, c'était aussi peut-être un problème pour les parents, qui manquent de disponibilité, qui doivent faire encore plus de choses à la demande des docteurs. Donc nous nous étions dit que ce serait bien de faire un questionnaire de stress pour savoir comment

3. FDA : Food and Drug Administration (USA).

les parents auraient vécu cette étude. Finalement, nous nous sommes rendu compte que c'était plutôt soulageant, et qu'il n'y avait pas d'augmentation du stress pour les parents. Au contraire, ils étaient assez contents parce que nous avions pris des enfants avec autisme assez jeune, qui commençaient le traitement [et cette activité venait donc en quelque sorte apaiser leur stress]. Souvent, quand il y a des annonces de diagnostic, les parents sont assez déroutés. Ils sont d'abord bien impactés par le diagnostic annoncé, et ils ne savent pas s'ils vont bien faire pour accompagner l'enfant. Donc nous avons compris *a posteriori* que d'avoir fourni un outil à la maison permettait d'avoir cette petite demi-heure pour jouer avec l'enfant (et le jeu de ce point de vue-là fonctionnait assez bien). C'était soulageant pour eux, car ils avaient l'impression d'être à la fois utiles à leurs enfants tout en jouant avec eux. Sur l'aspect utilité, nous n'avons pas encore démontré que ce jeu avait un intérêt en tant que tel. Mais dans la phase pilote, nous nous sommes posé la question de la surcharge d'impliquer les parents, et nous avons eu une réponse assez rassurante.

Difficultés institutionnelles

Mehdi Khamassi [29:19]

Quelles difficultés institutionnelles réglementaires ou juridiques tu as pu rencontrer, ou tes équipes, avec ce genre d'approches ?

David Cohen

Il y a trois difficultés différentes. Il y a une difficulté interne à mes équipes à appréhender cette question des nouvelles technologies. Après 10 ans nous n'avons plus ces difficultés-là. Mais au début, il y avait clairement de grands fantasmes collectifs. Par exemple, la première fois que j'ai parlé du robot AVS, tout le monde à l'hôpital se demandait s'il y aurait encore un jour des thérapeutes dans les services. Alors que ce n'est pas du tout ce dont nous parlons. Mais il y avait ce fantasme-là. La deuxième chose est qu'il y a dans le monde de la psychiatrie en France une certaine prudence par rapport aux données quantifiées ; il y a une sorte de charme du qualitatif. Le qualitatif peut être très utile sur certains aspects, mais le personnel au quotidien n'est pas toujours fan de compter, ni de faire des échelles. Pourtant, pour pouvoir faire de la recherche, nous avons quand même besoin de pouvoir numériser ce qu'on fait, et ainsi d'essayer de répondre à un certain nombre de questions avec des méthodes statistiques ou d'ingénierie. Donc on a toujours cette difficulté quand on fait de la recherche en psychiatrie. La troisième difficulté tourne autour du fantasme de la surveillance. En effet, en mettant en place des méthodes qui permettent de répondre à des questions (*i.e.,* voir ce qu'on n'arrive pas à voir et entendre ce qu'on n'arrive pas entendre), on peut glisser très vite vers « on surveille tout le temps ». Du coup, ça devient compliqué. Mais finalement concernant ces difficultés institutionnelles, dès que les gens voient les limites réelles du système, et surtout ce qu'impliquent concrètement ces petites expériences, ça se passe très bien.

Par contre, sur le plan administratif et juridique, pour moi la vraie difficulté me semble liée à une méconnaissance de la réalité de ce que nous faisons au sein des comités d'éthique. Souvent, il y a une confusion avec la question des *big data*, justement. Le fait que nous travaillons avec des ingénieurs sur des thématiques « d'une certaine modernité » peut donner l'impression que nous faisons forcément des partages de données. Pourtant nous faisons le plus souvent des expériences monosites de petite envergure en termes de nombre de sujets inclus. Nos expériences peuvent générer énormément

de données, donc nous faisons quand même du *big data*, parce que nous allons analyser 2000 signaux sociaux en même temps, par exemple. Mais ce ne sont pas du tout des données que nous partageons sur le Web. Et souvent on nous répond « comment ça se fait que vous n'ayez pas parlé de ça ? », ce qui amène à devoir donner tout un ensemble de justifications.

De plus, il nous faut à nous aussi prendre l'habitude des nouvelles réglementations, type RGPD, qui sont relativement récentes, dans la justification des protocoles. C'est aussi une difficulté, avec parfois des contraintes qui nous sont mises et qui rendent les expériences presque impossibles. Par exemple, le stockage des vidéos. Analyser les enregistrements vidéo des expériences peut être utile pour analyser le mouvement, les émotions faciales, la position du regard, etc. Nous avons souvent besoin des vidéos, car c'est le matériel sur lequel nous travaillons. Après nous pouvons aussi travailler sur des cartes de profondeur, sur les squelettes, mais disons la vidéo est le matériel sur lequel le mouvement et les émotions sont le plus faciles à reconnaître. Or on nous répond qu'on ne peut pas garder les vidéos parce que les enfants sont reconnaissables. Et ça, ce sont des retours qu'on a souvent. Et si ce sont des enfants qui ont des diagnostics, on risquerait de pouvoir obtenir leur diagnostic rien qu'à partir de la vidéo. Mais cette crainte vient de l'impression que ces vidéos sont partagées sur des serveurs et que n'importe qui peut les regarder. Alors que ce n'est pas plus partageable qu'un dossier médical qui est informatisé dans un hôpital, puisque notre serveur est à l'hôpital. Mais comme c'est de la vidéo, et qu'on a l'idée que ça peut se retrouver sur Internet, alors on nous met des limites. C'est une limite que je comprends quand il n'y a pas de serveur sécurisé. Mais cela pose un certain nombre de difficultés.

Par exemple, dans un projet précédent, le comité d'éthique avait gardé comme positions de n'enregistrer aucune vidéo. Cela n'a pas empêché le projet, car fort heureusement on avait déjà fait une expérience préliminaire où on savait grosso modo les *features*, c'est-à-dire les caractéristiques, les plus intéressantes. Pour ce projet, nous avions donc néanmoins dû faire le pari que les caractéristiques que nous avions trouvées chez les jeunes adultes allaient être les mêmes chez les adolescents, ce qui n'était pas forcément évident. En conséquence, nous avons extrait les caractéristiques pendant les enregistrements de manière automatisée (c'est un gros travail qui a pris six mois pour développer le système), sans garder la vidéo. Et du coup nous avons travaillé sur ces données-là et avons ainsi pu confirmer nos résultats chez le jeune adulte. Mais nous n'avons pas pu vérifier s'il y avait de nouvelles caractéristiques plus intéressantes chez les jeunes adolescents, puisque nous ne les avons pas enregistrées. Voici un exemple de limites que l'on peut rencontrer, et qui limitent certaines questions de recherche. Je trouve que c'était une prudence qui était exagérée par rapport à la nature du projet.

Risques de remplacement du soignant par un robot ?

Mehdi Khamassi

Alors j'aimerais bien qu'on revienne sur cette question du fantasme du remplacement du thérapeute ou du psychologue par le robot. Ce que je comprends de ce que tu décris, c'est qu'il y a déjà des limites techniques sur ce que les robots peuvent faire s'ils sont tout seuls. Il y a aussi le fait que pour l'instant il s'agit d'un outil, qui va venir apporter d'autres choses, par exemple détecter d'autres signaux qui ne seront pas détectables à l'œil nu, qui va pouvoir avoir un comportement bien contrôlé pour mettre en œuvre une certaine expérience et mesurer certaines choses. Mais au fond, j'ai déjà entendu dans certains projets européens qui utilisent des robots dans le cadre de l'autisme, l'argument comme quoi, avec le vieillissement de la population en Europe, au fur et à mesure des décennies on

va avoir moins de personnel soignant par personne, et qu'alors avoir des robots peut constituer un bien et devenir un objectif en soi. On se demande alors dans quelle mesure, même si à l'heure actuelle ce n'est pas l'objectif que l'on poursuit, si on ne va pas finir par y arriver. Est-ce que c'est un problème ? Est-ce qu'on peut faire en sorte qu'on y aille différemment ? Ou bien c'est juste un outil qui vient s'ajouter à ce qu'on fait actuellement ? Qu'est-ce que t'en penses ?

David Cohen

Je ne suis pas dans une position qui puisse faire changer les décideurs pour les faire rêver sur ce que sera le monde de demain. Et je comprends très bien, comme cela est mis à l'honneur dans le cinéma depuis des années, qu'avoir la possibilité de robots humanoïdes avec de vrais processus de pensée, de réflexion, d'accompagnement peut être tentant. Mais concrètement nous n'en sommes pas là, et nous sommes encore loin du compte.

Néanmoins pour revenir à la question du personnel soignant et du vieillissement, on peut imaginer qu'un certain nombre de tâches de vérification pour les personnes âgées peuvent être automatisées. Ceci ne requiert pas forcément de la robotique, parce qu'il n'y a pas forcément besoin d'avoir le robot au milieu ; il suffit parfois simplement d'installer une caméra, des systèmes de surveillance de potentiels dangers (du feu, de l'électricité, et de toutes les choses dangereuses quand on peut avoir des problèmes de sénilité). J'imagine que c'est déjà le cas sur des sites expérimentaux, mais je pense qu'on n'est pas loin d'un certain nombre de dispositifs qui pourraient être déployés à grande échelle et qui ne guériront pas la maladie d'Alzheimer, par exemple. Mais ce n'est pas impossible que ce soit des systèmes qui permettent à des personnes en cours de maladie d'Alzheimer de gagner un an ou deux ans de plus à leur domicile.

Dans les problématiques qui sont les nôtres, comme l'autisme, ce qui est intéressant avec les machines ou même les avatars – c'est-à-dire une autre forme de robots, mais virtuels cette fois-ci –, c'est que les enfants peuvent y trouver un intérêt particulier et stimulant. Il faut d'abord souligner la grande différence aujourd'hui entre les avatars et les robots : les premiers ont bénéficié des recherches dans le domaine des jeux, et on a de ce fait une qualité de mouvement, une subtilité dans le comportement, qui sont assez incroyables. C'est quelque chose qu'on n'a pas forcément en robotique. Mais en même temps, de nombreuses équipes ont fait des expériences pour montrer que notre rapport à l'avatar n'est pas tout à fait le même que notre rapport au robot, car le robot a une présence incarnée. La présence incarnée est un plus pour de nombreuses choses. Ce qui est intéressant en tout cas avec les enfants autistes, et les enfants en général, c'est qu'ils ont une tendance, voire une faculté, à l'animisme, comme l'écrivait Piaget dans *Le développement de l'enfant*. Ils vont donner de la vie même à la machine. C'est quelque chose de très intéressant, car le rapport à la machine n'est pas celui de l'adulte. Et qui plus est, pour l'enfant autiste, il y a un côté assez rassurant du robot, parce que le robot a ses limites, et finalement fait assez peu de choses. C'est relativement répétitif, mais c'est justement ce qui rassure l'enfant autiste. Donc on a ici une conjonction qui est assez heureuse. En contraste, les enfants normaux ont une très grande curiosité vis-à-vis des robots, mais si le robot est très répétitif, il finit par les ennuyer.

Je parlais du robot AVS : si on arrive à développer nos dispositifs de sorte qu'on arrive à une capacité d'interactions qui dure un certain temps, à savoir environ une heure (tel est notre objectif), alors cela permettrait de travailler avec un enfant une dimension spécifique. C'est toujours un gain par rapport aux soins. Aujourd'hui, dans l'autisme, un des grands résultats des recherches cliniques de ces dernières années est d'avoir montré que plus on prend en charge tôt un enfant autiste, mieux est leur devenir. Et surtout plus on les prend en charge intensément, mieux c'est. Tous les programmes qui prennent en charge au-dessus de 25 heures par semaine (ce qui est beaucoup) ont de meilleurs

résultats que les programmes où on a 2/3 heures par semaine de soin. Si le jeu sérieux ou le robot peut dans ces 20 heures en occuper 3 ou 4, c'est aussi ça de moins sur les dépenses de santé.

Un autre élément est qu'au-delà de la disponibilité interpersonnelle, il y a aussi le coût. Aujourd'hui le coût des dépenses de santé mentale est énorme. Il est énorme parce qu'on a des problématiques qui sont souvent des problématiques de « besogneux », c'est-à-dire qu'il y a plein de soins en psychiatrie qui doivent se faire en répétition. Les progrès sont lents. En effet, on aimerait avoir, comme dans d'autres pathologies, des médicaments qui guérissent en une semaine. Mais malheureusement dans toutes les chroniques en psychiatrie on n'a pas de traitement de ce type-là aujourd'hui. Et donc on est dans des prises en charge de « besogneux » qui ont une efficacité lente. Et donc ils coûtent très cher.

Le robot comme médiateur et facilitateur des interactions humain-humain

Mehdi Khamassi [45:20]

Quand tu décris ce lien entre l'enfant avec autisme et le robot, il y a un discours que j'ai entendu plusieurs fois de la part de collaborateurs que nous avons en commun[4]. Ceux-ci disent qu'ils ont observé un certain nombre de fois des enfants avec autisme préférer interagir avec des robots qu'avec des humains, parce que les premiers sont davantage prédictibles et plus limités. Dans certains cas, les enfants vont se mettre à regarder le robot dans les yeux, avoir des interactions peut-être davantage facilitées. Puis les chercheurs ont observé dans un deuxième temps que ça pouvait se répercuter dans les interactions de l'enfant avec les humains, avec ses parents. Il y a donc, une sorte de rôle de médiation par le robot, d'intermédiaire entre les enfants et les humains, pour aider *in fine* les enfants autistes à mieux interagir avec d'autres humains. Je me demande à quel point c'est encore quelque chose d'anecdotique, de ponctuel, ou si ce sont des choses qui sont vraiment démontrées et observées de manière répétée ?

David Cohen

Pour l'instant, ça reste anecdotique et ponctuel. Néanmoins, c'est une observation que je partage. C'est anecdotique et ponctuel, car aujourd'hui il y a peu d'équipes qui travaillent réellement en recherche en robotique sur le développement de l'enfant. Après, et c'est vrai en particulier avec le robot NAO, il y a eu un moment donné un travail assez important de la société[5] qui vendait le NAO pour en introduire dans des lieux cliniques qui ne faisaient pas forcément des recherches. Ce qui est dommage, c'est qu'on a assez peu d'études qui donnent un réel retour. Il y a un intérêt pour ces équipes d'avoir introduit un robot comme NAO, avec ses spécificités, dans leur organisation en termes de soin. Cela fait qu'on a des descriptions anecdotiques sur les résultats obtenus. Par exemple, des résultats sur un groupe de danse où on voit des enfants qui dansent avec NAO. Et on

4. NdE : les travaux pionniers de la chercheuse en robotique sociale Kerstin Dautenhahn, et de son collaborateur Ben Robins, à l'université de Hertfordshire au Royaume-Uni, ont montré que des enfants avec autisme se sentent plus à l'aise dans l'interaction avec un robot plutôt qu'avec un humain, car le robot a un comportement plus simple et prévisible, avec moins d'expressions faciales et donc moins de subtilités et de complexités dans l'expression des émotions. Dans certaines *success-stories*, cela a même amené des enfants autistes à finir par regarder le robot dans les yeux quand ils ne le faisaient jamais pour les humains, puis à finir par regarder leurs propres parents dans les yeux. Voir par exemple : Robins B, Dautenhahn K, Dubowski J (2006) Does appearance matter in the interaction of children with autism with a humanoid robot? Interact Stud 7(3):479–512.
5. NdE : Aldebaran Robotics, basée à Paris, puis devenue Softbank Robotics Europe.

voit clairement que les enfants sont amusés d'avoir le robot au milieu. Mais il n'y a pas de recherche qualitative, à ma connaissance, pour décrire assez précisément de manière globale ce qui se passe avec cet usage (qui est un usage presque commercial). Quand on aura une étude de ce type-là (car les robots se démocratisent), je pense qu'on aura peut-être des réponses. Ce ne seront pas des réponses *evidence-based*, comme en clinique. Mais ce sera tout de même des réponses sur le plan de l›usage social, au même titre qu'avec les jeux sérieux ou les jeux vidéo, qui sont maintenant une banalité. Je pense que d'ici 10-20 ans, tout le monde aura un robot chez soi. Déjà les robots qui font le ménage sont plus répandus à domicile. Je pense qu'il y aura alors des retours d'expériences d'usages qui pourront éclairer l'intérêt naturaliste que peut avoir ce type d'objets avec des enfants avec autisme ou avec des troubles du neuro-développement.

Dans l'autisme, les progrès sont souvent spécifiques et peu généralisables

Mehdi Khamassi

Même si ses observations restent anecdotiques, on a envie que ça soit plus systématiquement démontré et quantifié pour que certains des éléments scientifiques puissent venir à l'appui de ce type de progrès.

David Cohen

Je suis tout à fait d'accord avec ta remarque. En même temps, la question de la généralisation, c'est-à-dire de l'extension de ce qu'on peut voir pour un enfant en termes d'amélioration dans le cadre d'un protocole spécifique, en l'occurrence un protocole spécifique avec un robot ou avec un jeu de sérieux, c'est un problème dans l'autisme en tant que tel. C'est-à-dire qu'à chaque fois qu'on entraîne un enfant à quelque chose, on va dire qu'il va un peu progresser sur ce quelque chose spécifique, parce qu'il y a une plasticité cérébrale et des aspects développementaux. C'est un progrès dans l'entraînement. Mais en fait dans l'autisme cette question de la généralisation est vraie pour tous les traitements possibles. C'est pour ça qu'il est intéressant d'avoir une vision assez globale des activités de soins. Et c'est pour ça que c'est intéressant d'avoir un partenariat intégré. Car si on est ingénieur on s'intéresse et on connaît bien le robot, ou le jeu sérieux (l'objet de son travail). Mais si on est en discussion avec des cliniciens qui travaillent sur les formes de réponses attendues avec d'autres types d'entraînements, à ce moment-là on voit que la question de la généralisation est vraiment compliquée à démontrer. Et encore une fois, ceci ne sera, à mon avis, pas spécifique de la robotique (si jamais un jour on atteint ce stade-là de la démonstration), mais ça sera plutôt spécifique de l'autisme. Parce que dans l'autisme, une des grandes difficultés est justement cette faculté à pouvoir généraliser, qui est une faculté développementale, et qui chez l'enfant normal est vraiment époustouflante.

Florian Forestier

J'ai une question d'ordre épistémologique, qui part de mon point de vue et de mon intérêt pour les nouvelles technologies dans le domaine de l'autisme. J'ai participé à la préparation de la stratégie nationale autisme dans le groupe recherche, et je suis dans plusieurs comités de suivi de thèses sur ces sujets, notamment un avec Mohamed Chetouani, à l'ISIR également. Il y avait une première

question qui était double dans le cas de l'autisme : 1) la possibilité d'avoir une observation et d'avoir un contexte d'observation qui ne biaisent pas totalement les résultats ; l'idée que, dans le cas d'enfants et de personnes autistes, les conditions de laboratoire sont beaucoup plus perturbantes, et donc beaucoup plus biaisées que dans un cas normal, et donc que les nouvelles technologies pouvaient être très intéressantes pour ça. Ceci constitue le premier aspect. Sur ça, il y avait un certain consensus, mais, derrière ça, il y a un désaccord sous-jacent sur la manière de faire. Ceci est dû au fait que certains chercheurs importent le modèle des *living labs* pour répondre à ça ; alors que ce modèle vient plutôt de la gériatrie, avec l'idée d'espaces qui vont être connectés. Pour ma part, je fais plutôt partie de la filiation de ceux qui disaient que le cas de l'autisme invite à remettre en cause la dichotomie un peu trop massive entre conditions écologiques et conditions de laboratoire, pour penser vraiment la question de la contextualité.

David Cohen

lors, je vais prendre les questions une à une. En fait, je comprends ce type de questionnements. Je suis plutôt d'accord sur le fait que dans l'environnement laboratoire, dans l'environnement expérimental (car ce n'est pas l'environnement naturaliste ou écologique de la maison ou de l'école), quand on peut faire des observations, qu'elles soient via de nouvelles technologies ou via des humains (parce qu'on peut aussi faire des observations à l'école avec des observateurs et avec d'autres méthodologies), on trouve des résultats effectivement différents. Le plus souvent ces résultats sont de meilleure qualité pour les enfants autistes, parce qu'ils connaissent déjà leur environnement écologique. Pour autant, beaucoup d'expériences dites de laboratoire (c'est là que quand on n'est pas totalement dans les actions de terrain, on fait parfois des dichotomies qui sont un peu artificielles, parce qu'en fait beaucoup des expériences qui sont faites avec les nouvelles technologies dans l'autisme sont en fait de deux ordres : soit ce sont des expériences type *big data*, où on récupère des données en milieu écologique, mais avec des outils qui sont des outils de bas grades au niveau de l'ingénierie ; ça va par exemple être les mouvements avec le téléphone, ça va être comment il dort le jour et la nuit, ça va être des questions de cet ordre, des questions qui peuvent avoir un intérêt, encore une fois mais qui sont basées sur des choses relativement simples au plan technologique) ; et puis, quand on parle de robotique, ici on parle de quelque chose d'un peu différent, qu'on n'arrive pas à mettre facilement à la maison. Par contre, ces recherches sont menées souvent par des petites équipes de recherche, et les robots sont intégrés aux milieux d'où viennent les enfants.

Par exemple, dans notre équipe, toutes les expériences qui sont faites dans le cadre d'un hôpital de jour, où l'on a des enfants qui viennent régulièrement, font en sorte que ces enfants ont l'habitude des ingénieurs de la salle expérimentale, et même de certains robots que nous utilisons (c'est presque devenu des copains). Je me souviens que, sur un projet, un enfant m'avait interpellé et m'avez demandé « Et alors, quand est-ce qu'on voit NAO ? » Donc ce côté n'est pas si dichotomique que ça, même si je comprends que l'on se pose la question de cette manière. Après, ce qu'on a fait en gériatrie était aussi pour répondre à des questions spécifiques de surveillance. Ce ne sont pas du tout les mêmes questions que dans l'autisme. Or, pour moi, si on veut faire de la recherche qui n'est pas du gadget, il faut se demander : est-ce qu'avoir une caméra, un système intégré, un robot au milieu, éventuellement des données qui sont fournies par le téléphone, donc par des capteurs, est-ce que tout ceci nous procure réellement une architecture qui nous permet d›attendre quelque chose de cette captation ? C'est une question à laquelle on doit répondre. Souvent, les projets intégrés de cet ordre-là sont des projets où on veut capter le maximum d'informations, et après faire du *machine learning* dessus pour se poser des questions. C'est-à-dire qu'il n'y a pas réellement d'*a priori*. Pour ma part, je ne travaille pas tellement sur ce type de dynamiques. Quand nous faisons des efforts pour capter les informations

(c'est un énorme boulot de capter l'information correctement, et c'est rarement valorisé dans les projets), et bien je préfère que nous ayons quand même une question de recherche, et que nous anticipions la raison de capter cette information de cette manière-là. Et c'est pour ça que des fois nous faisons des études exploratoires pour aller plus loin. Finalement, le problème est que les choses sont souvent présentées comme des réponses à des problèmes complexes, alors que le raisonnement qui est derrière n'est pas réellement explicité.

Florian Forestier

Merci beaucoup. C'était justement mes deux points, car je suis d'accord avec vous. J'ai plutôt cette position-là, à la fois sur la question de ne pas faire une dichotomie trop forte, et sur la question de ne pas avoir une hypothèse de neutralité qui n'est finalement pas si neutre que ça, mais plutôt de créer des contextes, voire dans le cas de l'autisme, d'essayer de susciter des contextes qui vont créer des stimulations qui ne sont pas là dans l'environnement habituel. Et c'est souvent ça qui me semble manquer dans les projets actuels.

Intelligence collective sur les partenariats publics-privés incluant des robots dans la clinique

Mehdi Khamassi

La dernière question que je veux te poser est relative à l'intelligence collective, au niveau méta dans notre société, avec nos institutions, nos comités d'éthique, la façon de gérer les partenariats public-privé : est ce qu'il te semble qu'on est dans un processus d'intelligence collective sur la façon de gérer l'intégration de technologies dans la société telles que les robots ou l'IA ? Ou y a-t-il des choses qui manquent selon toi, des choses qui sont mal gérées ou qui vont trop vite, et qu'on n'a pas les bons garde-fous ? Comment vois-tu les choses ?

David Cohen

Je ne crois pas qu'on soit dans un climat d'intelligence collective. Pour moi, l'intelligence collective, c'est comment l'humain peut trouver une forme d'équilibre entre les intérêts individuels et les intérêts collectifs, les intérêts à court terme et les intérêts à long terme. Quand on étudie, ne serait-ce qu'un peu, l'Évolution, on sait que c'est une règle : l'Évolution tire toujours vers la survie dans une opposition entre effets à court terme et effets à long terme. Il y a alors des compromis qui sont trouvés en fonction de la complexité des systèmes à un moment donné. Et par ailleurs la même chose se produit entre l'individuel et le collectif. Cette règle s'applique aussi avec ce qu'il se passe sur les nouvelles technologies. Et elle s'applique aussi au contexte de l'autisme. On ne peut pas généraliser au monde entier, mais ceci est vrai ne serait-ce que si on prend le contexte français. Pour ce qui est des nouvelles technologies, le contexte français est coincé dans des questionnements où on se demande si le collectif ne prend pas le pas sur l'individuel. Ceci est en particulier dû au fait que la puissance commerciale d'un nombre de grands groupes est telle qu'ils sont devenus quasiment plus puissants que les institutions qui nous gouvernent. Ceci en dit long sur une forme de dilution de la responsabilité individuelle, et même de la responsabilité des institutions. Alors, même si les institutions font des déclarations d'intention pour dire « on veut contrôler, on veut vérifier ceci ou cela »,

elles se retrouvent en difficulté parce qu'il y a des enjeux financiers, des enjeux de rapports de force entre pays. Tout ceci fait que c'est très difficile de trouver une forme de consensus collectif, qui serait une forme de sagesse collective. Donc au niveau des nouvelles technologies, il y a cette dimension-là.

Puis la deuxième dimension est celle que nous avons évoquée au niveau des cas particuliers, avec tous les fantasmes que ça suscite. Moi qui ai travaillé sur un sujet totalement différent, celui de la biologie de la reproduction, j'y ai vu la question être posée : dans quelle mesure ces nouvelles biologies de la reproduction vont ou non impacter le développement des enfants ? Nous avions alors monté un groupe de réflexion multidisciplinaire, à la fois clinique et éthique, pendant plusieurs années. Et j'ai trouvé extrêmement intéressant de voir dans ce groupe à quel point toutes ces modifications que l›on voit dans la génétique et la biologie de la reproduction (donc des biotechnologies) reviennent toujours à une forme de parti pris de technoprophètes qui pensent que, parce qu'on a découvert certaines choses, on aura la réponse à tout. A contrario, il y a parfois les grands sceptiques, qui ont l'impression que, parce qu'on a découvert quelque chose, c'est la fin du monde, on ne fera plus de famille judéo-chrétienne avec une filiation digne de ce nom, et c'est donc la mort annoncée de l'Humanité.

Finalement, ce genre de tendances se retrouvent aussi pour les nouvelles technologies telles que la robotique ou les *big data* : vous avez ceux qui croient que c'est l'avenir du monde, et ceux qui sont dans un truc de persécution et de théories conspirationnistes, pensant que c'est la fin du monde. Certains disent les théories conspirationnistes sont contemporaines. Mais c'est totalement faux. Des historiens ont montré qu'il y a déjà eu des moments conspirationnistes, certes pas à la même échelle du fait de la communication, mais il y en a eu, notamment au Moyen-Âge. On en trouve des traces et il y a des documents qui nous permettent d'en rendre compte. Donc ça a existé, d'autant plus que dans l'Histoire humaine il a fallu gérer en très peu de temps ces extraordinaires modifications qu'était la formation des villes. Car quand on était chasseur-cueilleur et qu'on allait se balader dans les grandes forêts, certes il y avait des groupes concurrents, mais le collectif n'avait pas la même valeur, et la taille du groupe d'affiliation n'avait pas du tout la même valeur que quand on est 10 millions sur 20 ou 100 km². Forcément, tout ça crée des modifications majeures. Et je pense que ce sont ces tensions qu'on voit à l'œuvre pour ces questions concernant les nouvelles technologies, qu'elles soient biologiques ou numériques.

Par contre, le champ de l'autisme est un autre contexte, qui a ses propres lignes de force. Il y a un courant assez fort en médecine, qui est venu des courants gays et lesbiens des années 1960. C'est un courant qui a été revendicatif au niveau de la manière dont la psychiatrie envisageait à une certaine époque la question de l'homosexualité, ou même de la dysphorie de genre ou de la transidentité. Et donc c'est une sociologie de combat. À un moment donné, les associations se sont opposées au point de vue médical, qui était un point de vue très paternaliste. Et c'est grâce à cette opposition qu'elles ont gagné en légitimité et ont pu modifier le système. Et d'une certaine façon, cette manière combative dans les associations de représentants de patients a infiltré tout le champ de la médecine. Comme les sociologues le racontent très bien, les premiers mouvements puissants ont été dans ce domaine-là.

Dans l'autisme, nous sommes aujourd'hui dans un climat de cet ordre-là. Il y a un certain nombre d'associations qui sont très revendicatrices, plutôt sur un mode « combat ». Et aujourd'hui, elles font peur aux politiques. On est donc dans un contexte où il faut tenir compte de cette dimension-là quand on discute de quoi que ce soit. Si on ne comprend pas la complexité de ces enjeux, qui sont des enjeux macro, on ne peut pas répondre aux enjeux micro. Par exemple, aujourd'hui, il se trouve que je suis membre de la commission nationale de psychiatrie pour la pédopsychiatrie (mise en place par la DGOS à la demande du ministère de la Santé). Et, clairement, une des questions posées dans l'organisation des soins est : comment peut-on réconcilier les parcours sanitaires et les parcours qui sont dans le médico-social ? En effet, quand on a un handicap chronique, en France, très souvent, on est entre le sanitaire et le médico-social. Cela génère des aller-retour. Et le médico-social est, au-

jourd'hui, essentiellement géré par des associations de familles et de représentants de parents. Tandis que le médical est géré, pour l'autisme, par la psychiatrie de l'enfant et de l'adolescent, en tout cas en très grande majorité. Et il y a 20 ans d'opposition entre les deux, puisqu'il y a eu un gros combat. Et d'ailleurs j'imagine que vous avez ce type de combats dans les réunions du ministère, même si ces réunions concernent la recherche et les technologies. Mais je pense qu'ils sont obligés d'en discuter.

Florian Forestier

Ceci est encore plus vrai sur le programme de recherche en sciences humaines. La conflictualité est à la fois un objet d'étude et un paramètre général.

Mehdi Khamassi

Il est temps de conclure. Merci beaucoup pour ton temps, David.

David Cohen

Merci !

Vers une interdiction du traitement automatique des émotions ?

Célia Zolynski

CÉLIA ZOLYNSKI

Célia Zolynski est professeur agrégée de droit privé à l'Ecole de droit de la Sorbonne de l'Université Paris 1 Panthéon-Sorbonne où elle dirige le Master 2 Droit de la création et du numérique et codirige le Département de recherche en droit de l'immatériel de la Sorbonne (IR JS-DreDis). Membre du Comité national pilote d'éthique du numérique (CNPEN), elle est en outre personnalité qualifiée au sein de la Commission consultative nationale des droits de l'Homme (CNCDH) et du Conseil supérieur de la propriété littéraire et artistique (CSPLA).

Ce texte, issu de la séance du séminaire interne TESaCo du 4 Avril 2023, a été mis à jour par l'auteur en janvier 2024 pour tenir compte des dernières avancées du Règlement sur l'IA de l'Union européenne.

I. Le traitement automatique des émotions : l'aube d'une ère nouvelle ?

Ma présentation portera sur l'analyse du cadre juridique concernant le traitement automatique des émotions de l'humain dans le cadre de ses interactions avec la machine. Il s'agit d'un sujet que j'ai envisagé dans un article récent coécrit avec Judith Rochfeld[1], mais aussi dans des contextes différents, notamment au sein du Comité national pilote d'éthique du numérique *(Avis n°3, Agents conversationnels : enjeux d'éthiques,* sept. 2021) et de la Commission Nationale Consultative des Droits de l'Homme (*Avis A-2022-6 relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux*, 2022).

Ces discussions se poursuivent désormais à l'occasion des négociations en cours concernant la proposition de Règlement sur l'intelligence artificielle de l'Union européenne, avec des implications notables dans le contexte du déploiement massif d'agents conversationnels.

Le point de départ de notre réflexion résidait dans l'observation d'annonces significatives concernant le déploiement de solutions, en par-

1. C. Zolynski, J. Rochfeld, « Valorisation des émotions : quel encadrement pour le capitalisme cognitif ? », *Entre art et technique : les dynamiques du droit. Mélanges en l'honneur de P. Sirinelli*, Dalloz, 2022, pp. 749-770.

ticulier dans le domaine de la reconnaissance émotionnelle, accompagnées de données chiffrées. Un article publié dans Nature annonçait ainsi une estimation de l'ordre de 37 milliards de dollars pour 2026 concernant l'industrie de la reconnaissance des émotions[22].

Actuellement, de nombreuses promesses émanent du secteur industriel en matière de détection et d'interprétation des manifestations de diverses émotions humaines, y compris les plus intimes. Ces manifestations incluent l'impatience, la crainte, le désir, exprimés à travers des signaux tels que le clignement des yeux, les variations de la voix, de l'écriture, du rythme cardiaque ou encore le léger affaissement des mâchoires, entre autres. Par ailleurs, on observe l'émergence - ou du moins l'annonce - de plusieurs projets, comme celui d'Amazon en août 2020 avec le lancement de son bracelet connecté Halo. Ce dispositif permettait de traiter l'image en 3D du corps et les tonalités émotionnelles de la voix de l'utilisateur afin d'évaluer ses habitudes néfastes et de l'encourager à adopter des comportements plus bénéfiques. Parmi les exemples, on peut également évoquer celui de la société Muvraline proposant des outils de reconnaissance émotionnelle basés sur l'image. Ces outils sont conçus pour détecter automatiquement l'état émotionnel de personnes, telles que celles manifestant de la nervosité, de l'agressivité, ou de la crainte, par exemple lors d'un embarquement en avion ou dans des espaces publics. La société affirme également être en mesure d'évaluer de manière non invasive et fiable l'attrait des produits, ainsi que la capacité d'un contenu publicitaire à susciter efficacement l'attention et l'intérêt des spectateurs.

Pendant la période de confinement à la suite de la pandémie de Covid19, la société 4Little Trees s'est distinguée en proposant une méthode visant à évaluer l'attention des élèves en classe. Selon ses concepteurs, cette approche devrait permettre aux enseignants d'ajuster leurs pratiques pédagogiques et d'évaluer de manière plus précise les motivations des apprenants ; une méthode déjà utilisée notamment en Chine. Nous observons également une expansion significative de l'utilisation de systèmes de reconnaissance émotionnelle dans le domaine des ressources humaines. Ces dispositifs sont particulièrement employés pour évaluer les candidats lors du processus d'embauche ainsi que le comportement des employés dans l'exercice de leurs fonctions, notamment lorsqu'ils sont confrontés à des situations de stress ou de risque de *burnout*.

On peut en outre inclure dans cet ensemble tous les projets de métavers – et en particulier celui annoncé par la société Meta bien que les promesses initiales soient moins convaincantes aujourd'hui qu'au moment de son lancement. Plusieurs de ces projets visent à capturer de manière plus précise l'éventail complet des émotions des utilisateurs de ces nouveaux systèmes, en utilisant des casques de réalité virtuelle et autres dispositifs haptiques. L'objectif est de capter de nouvelles données d'interactions et d'ajuster en temps réel l'environnement virtuel dans lequel l'avatar humain évoluera.

S'il est vrai que l'analyse des émotions n'est pas une nouveauté, il nous a semblé qu'on était potentiellement à l'aube d'une ère nouvelle, celle d'une *transparence émotionnelle* qui pourrait comporter des risques de violation de l'intimité individuelle et de l'autonomie humaine sans précédent.

Il faut préciser qu'un certain nombre de ces craintes sont toutefois relativisées en raison de critiques formulées à l'heure actuelle en raison de l'absence de preuves scientifiques solides concernant la lecture potentielle et l'interprétation des émotions ou de l'état d'esprit d'une personne à partir de son visage, de l'analyse de sa démarche, de son rythme cardiaque ou de la tonalité de sa voix. Toutes ces analyses sont encore de l'ordre du prospectif et dépendent de leur caractère éminemment contextuel et culturel. Pour cette raison, la fiabilité de la reconnaissance émotionnelle demeure contestable, pour le moins en l'état actuel des techniques.

2. K. Crawford, « Time to regulate AI that interprets human emotions », *Nature*, vol. 592, 8 avril 2021, p. 167.

Néanmoins, cette technologie et ses possibles développements suscitent, d'ores et déjà, un certain nombre de questions qu'il est essentiel d'examiner dès à présent.

Pour ce faire, il convient de présenter une première typologie de ces pratiques, certes encore incomplète mais qui a pour vocation de faire émerger un certain nombre de questionnements concernant, en particulier, diverses utilisations du traitement des émotions à des fins commerciales - que nous avons qualifié de *capitalisme mental* ; nous mettrons ici de côté le traitement des émotions dans le domaine de la santé, y compris les traitements issus par exemple de l'imagerie cérébrale, qui peuvent être considérés comme des dispositifs distincts. Il s'agira d'exposer ensuite l'état actuel du droit applicable pour en identifier les insuffisances ce qui conduira, enfin, à formuler diverses propositions pour faire évoluer le cadre juridique.

II. De la captologie à l' « affective computing »

Il existe différents niveaux de traitement des émotions que l'on peut commencer à identifier. Un premier niveau consiste, à partir de l'analyse d'émotions préexistantes, à influencer le comportement de la personne en exploitant ses informations, par différentes techniques comme la *captologie* ou la possibilité de manipuler le comportement de l'utilisateur de l'interface avec la machine. Le deuxième niveau résulte de la possibilité de susciter de nouvelles émotions, notamment par le biais de techniques qualifiées d'*affective computing*, ou plus généralement d'utilisation d'agents conversationnels.

Évoquons tout d'abord les techniques de captologie[33], déjà cartographiées par la CNIL[44], qui font l'objet d'une attention importante de la part d'un grand nombre d'institutions publiques, même au niveau international, comme l'OCDE. Dans ce cadre, une préoccupation importante émerge, tant par rapport à ce que l'on peut qualifier de *bad nudge*, une technique qui incite l'utilisateur à prendre une décision allant à l'encontre de ses intérêts, qu'en ce qui concerne les *bad sludge*, visant à empêcher l'utilisateur d'agir selon son intérêt, par exemple en créant une friction, une difficulté artificielle, dans le but d'entraver sa liberté de choix. Ces techniques pourraient prendre une tout autre ampleur avec le développement de ce qu'on pourrait appeler les *augmented dark patterns*.

Il est en effet possible d'imaginer, en faisant notamment référence aux rapports de l'OCDE cités plus haut ainsi qu'aux travaux de l'Autorité de régulation des marchés et des consommateurs du Royaume-Uni, que les interfaces de choix pourraient être adaptées en temps réel par un traitement algorithmique de grandes masses de données et des procédures d'A/B testing qui pourraient conduire à personnaliser les contenus en se basant sur les biais cognitifs de la personne et adapter les interfaces en conséquence. De ces possibles scénarios, une situation préoccupante émerge, puisque ces technologies pourraient amplifier, voire démultiplier dans un certain nombre de situations, le tas de vulnérabilité de l'humain interagissant avec de l'interface.

On peut citer également de nouvelles formes de publicité basée sur les émotions (ou *emotion driven advertising*) qui viseraient à calculer les émotions afin de les restituer sous une forme d'information comportementale. Naturellement, cette possibilité pourrait être utilement mobilisée dans le secteur du marketing pour offrir des offres particulièrement adaptées à la personne. Ce type d'utilisation rejoint d'ailleurs certains projets de métavers pour adapter l'environnement dans lequel la personne évoluera au travers de son avatar. De ces possibles scénarios, une situation préoccupante émerge puisque ces pratiques pourraient amplifier,

3. Ces techniques ont fait l'objet de discussions au sein de TESaCo, dans le cadre de la préparation du livre co-écrit par Stefana Broadbent, Mehdi Khamassi, Florian Forestier et Célia Zolynski, intitulé *Pour une nouvelle culture de l'attention*, à paraître chez Odile Jacob en mars 2024.
4. *La forme des choix. Données personnelles, design et frictions désirables*, Cahiers IP n°06, 2019.

voire démultiplier dans un certain nombre de cas, les situations de vulnérabilité de l'utilisateur de divers services numériques.

Ces problématiques semblent encore plus prégnantes en ce qui concerne le recours à l'*affective computing* - ou informatique affective -, qui ne permet pas uniquement de traiter des émotions pour en déduire des informations, mais qui permet en outre de générer chez l'humain des nouvelles émotions par ses interactions avec la machine s'adaptant à son état émotionnel. Citons à titre d'exemple la start-up Emoshape qui propose de déterminer en temps réel les émotions des utilisateurs et de permettre aux robots et autres applications de répondre en simulant un état émotionnel en phase avec celui de l'utilisateur. L'objectif sous-jacent est de développer une stratégie de dialogue plus efficace et d'accroître ainsi l'engagement de l'utilisateur.

On peut alors envisager que les émotions humaines soient perturbées par la machine, notamment lors d'échanges avec un agent conversationnel. Un exemple presque caricatural, mais pas pour autant fictionnel, est celui du déploiement des «deadbots», ces agents conversationnels qui tirent parti des conversations enregistrées pendant la vie d'une personne pour prolonger d'une manière artificielle sa personnalité après son décès. Un «deadbot» peut ainsi continuer à interagir avec, par exemple, les proches du défunt. À cet égard, le Comité national pilote d'éthique du numérique a formulé diverses propositions pour protéger la personne dont les traces seraient exploitées après son décès. L'objectif est d'éviter que sa personnalité soit altérée ou manipulée par ces nouvelles formes de conversation, tout en considérant le désir d'être prolongé après la mort, et que ces pratiques portent atteinte à la dignité de la personne humaine. Par ailleurs, l'utilisation de «deadbots» peut soulever des questions concernant les risques de manipulation des proches en raison de cette relation particulière avec un agent artificiel prolongeant la personnalité du défunt. Au Japon, de nombreux projets sont en cours pour encadrer ces technologies. Bien que cela mérite une discussion distincte, il s'agit d'un exemple éclairant des risques associés à l'exploitation des données émotionnelles et à la simulation artificielle des émotions par les agents conversationnels.

Ces différents aspects sont notamment décrits dans les travaux de Laurence Devillers qui met aussi en lumière des risques d'une autre nature liés à la coadaptation entre l'humain et la machine, et en particulier des risques d'anthropomorphisation, c'est-à-dire le fait que l'humain projette des réactions affectives et probablement empathiques envers des entités artificielles[5]. Autant de problématiques qui nous avaient particulièrement préoccupés au sein du Comité national pilote d'éthique du numérique (Avis n°3, préc.).

Si l'on se limite aux considérations relatives aux émotions, on peut suivre l'analyse du psychiatre Serge Tisseron en ce qui concerne les enjeux découlant de ces nouvelles formes de conversations personnalisées. Tisseron décrit "une relation nouvelle de l'homme avec ces technologies numériques, de longues conversations suscitant un éventail d'émotions larges et variées, nourries par l'illusion d'une présence réelle, attentive et chaleureuse" pour terminer par un avertissement : "n'attendons pas et ne commettons pas la même erreur avec les machines parlantes qu'avec les téléphones mobiles" (S. Tisseron, *Vivre dans les nouveaux mondes virtuels, Concilier empathie et numérique*, Dunod, 2022, p. 188).

III. Les limites du droit actuel dans l'encadrement du capitalisme émotionnel

Compte tenu de l'essor de ce nouveau capitalisme émotionnel, il apparaît nécessaire d'analyser le cadre juridique applicable pour en déterminer les limites. Si l'on regarde le droit

5. L. Devillers, « Des interfaces traditionnelles humain-machine. Intelligence artificielle/intelligence humaine : manipulation et évaluation », Enjeux numériques, *Annales des mines*, déc. 2020, n°12, p. 78.

actuel, les limites que nous avons identifiées avec Judith Rochfeld se trouvent tout d'abord dans le droit des données à caractère personnel. Une deuxième limite peut résulter d'autres fondements afin de répondre aux risques résultant de la manipulation des émotions ce qui peut conduire jusqu'à préconiser une interdiction de ces pratiques, en particulier dans un contexte commercial.

En ce qui concerne le premier fondement, celui des données à caractère personnel, il est indéniable que les données émotionnelles, lorsqu'elles permettent de singulariser une personne et contiennent des informations spécifiques à une personne déterminée en raison de leur contenu, de leur finalité ou de leurs effets, soulèvent des préoccupations légitimes en matière de confidentialité et de protection des données. Il convient alors de déterminer si leur protection peut résulter de l'application du RGPD complété, en France, de la Loi informatique et libertés.

Si la définition compréhensive des données à caractère personnel permettra de retenir sans trop de difficulté cette qualification s'agissant des données émotionnelles, cela peut s'avérer plus délicat pour les données sensibles dont le traitement est interdit. Certes, cette qualification ne posera pas difficulté si les données révèlent l'état de santé de la personne, ce qui peut être le cas en cas de révélation par exemple d'un état dépressif, ou l'orientation sexuelle d'une personne ou encore en cas de traitement de données biométriques permettant d'identifier la personne. Cependant, ce régime n'est pas exhaustif en termes de protection étant donné les exceptions qui permettent le traitement de données sensibles dans certaines circonstances, à l'image de celle peu exigeante relative au consentement de la personne concernée. Compte tenu des enjeux, une telle qualification de données sensibles peut donc sembler insuffisante, d'où la proposition plus ferme que l'on porte de consacrer une interdiction des traitements sensibles des données émotionnelles, c'est-à-dire de tout traitement présentant des risques importants en termes d'intrusion dans l'intimité ou de limitation de l'autonomie de la personne, et ce quand bien même les données traitées ne seraient pas des données sensibles au sens de l'article 9 du RGPD[6].

Au-delà du traitement des données émotionnelles, d'autres limites peuvent être envisagées pour répondre au risque de manipulation des émotions de l'humain interagissant avec la machine.

Le droit des données à caractère personnel pourrait constituer une première barrière à ce type de manipulation, étant donné qu'il impose les principes de loyauté, de transparence du traitement, de *privacy by design*, ainsi que la garantie d'un consentement éclairé. Ces principes sont mis en avant par la CNIL, comme en témoignent les sanctions prononcées à l'encontre des interfaces de choix manipulatrices concernant par exemple les modalités d'acceptation des cookies, ainsi que par le Comité européen de la protection des données comme en attestent ses lignes directrices de mars 2022 sur les « Dark patterns » et les réseaux sociaux.

Outre le droit des données, d'autres fondements sont envisageables. Le recours aux interfaces trompeuses peut être sanctionné au titre des pratiques commerciales déloyales dès lors qu'un professionnel adopte un comportement non diligent influençant le comportement du consommateur. À cet égard, plusieurs propositions ont été formulées, y compris aux États-Unis, où une réforme du Consumer Privacy Act en Californie (CCPA) vise à sanctionner ce type de pratiques. En Europe, nous disposons aujourd'hui du Digital Services Act et du Digital Market Act, adoptés à l'automne 2022, qui établissent un principe d'interdiction des dark patterns.

Mais, compte tenu des éventuelles difficultés liées à la définition des interfaces trompeuses ou manipulatrices, il est également nécessaire de compléter ces premières approches en visant directement les systèmes algorithmiques. En ce sens, l'Europe a cherché à promouvoir une obligation de loyauté et de

6. Cela fait d'ailleurs écho aux principes de proportionnalité et de minimisation qui sont imposés par le RGPD.

transparence en ce qui concerne les systèmes d'intelligence artificielle interagissant avec des personnes physiques. Ainsi, l'article 52 de la proposition de Règlement sur l'intelligence artificielle publiée par la Commission européenne en avril 2021, prévoit que les personnes physiques devront être informées lorsqu'elles interagissent avec un agent virtuel. Une obligation de transparence est également exigée en cas de recours à un mécanisme de reconnaissance émotionnelle. On peut cependant considérer que ces mécanismes d'information ne sont pas suffisants pour prévenir les risques majeurs d'atteinte aux droits fondamentaux des personnes. Comme illustré par des exemples de «deadbots», même lorsque les personnes sont informées qu'elles interagissent avec une machine – ici simulant une personne décédée – cela peut néanmoins entraîner des troubles émotionnels importants. Cela souligne une possible insuffisance de la seule transmission de l'information.

IV. Vers une interdiction du traitement automatique des émotions ?

Pourquoi dès lors promouvoir l'interdiction de la reconnaissance émotionnelle ? Certains avancent cette idée en raison de l'absence de fiabilité de ces systèmes, comme mentionné précédemment. En raison de cette imprécision, les systèmes de traitement automatique des émotions pourraient être à l'origine de discriminations, comme l'a bien identifié le Défenseur des droits dans un rapport récent (Technologies biométriques : l'impératif respect des droits fondamentaux, juil. 2021). Ce rapport met en lumière l'utilisation de ces outils de reconnaissance émotionnelle en particulier dans le cadre de la gestion des ressources humaines. Comme la reconnaissance émotionnelle n'est pas encore fiable, ce traitement pourrait entraîner de faux positifs, des faux négatifs, et donc des risques de discrimination notamment lors des procédures d'embauche.

Il nous semble que la difficulté ne tient pas uniquement à la fiabilité des systèmes de reconnaissance émotionnelle compte tenu des risques de manipulation pouvant en résulter comme le relevait le Conseil de l'Europe (Conseil de l'Europe, Déclaration du Comité des ministres sur les capacités de manipulation des processus algorithmiques, fév. 2019). L'interdiction de la reconnaissance émotionnelle trouverait alors sa justification dans la nécessité de préserver l'autonomie cognitive de toute personne, son droit de se forger une opinion et de prendre des décisions indépendantes.

Une telle interdiction de principe a été portée par la CNCDH dans le prolongement des analyses du Contrôleur européen à la protection des données et du Comité européen de la protection des données. Des exceptions pourraient néanmoins être envisagées s'agissant de la recherche ou de la protection de la santé ou si l'on parvient à démontrer que ces systèmes renforcent l'autonomie de la personne plutôt que de la diminuer, par exemple, en favorisant des activités d'apprentissage pour des personnes souffrant de handicap.

On pourra alors regretter que la proposition de Règlement sur l'intelligence artificielle n'aille pas aussi loin. La Commission européenne avait fait le choix de restreindre l'interdiction à l'utilisation de systèmes d'IA dans le but d'influencer de manière subliminale le comportement d'une personne en vue de lui causer ou de causer à un tiers un dommage, ou encore les systèmes exploitant la vulnérabilité d'un groupe de personnes. Le Parlement européen a pour sa part proposé, dans son accord de compromis de juin 2023, de prohiber les systèmes de reconnaissance des émotions utilisés pour les services répressifs, la gestion du contrôle aux frontières, sur les lieux de travail et dans les établissements d'enseignement. Outre ces quelques cas d'interdiction, il préconise d'encadrer le déploiement des systèmes destinés à être utilisés pour effectuer des déductions sur les caractéristiques personnelles des personnes physiques sur la base de données biométriques ou fondées sur la biométrie,

y compris les systèmes de reconnaissance des émotions et s'ils présentent un risque important de préjudice pour la santé, la sécurité ou les droits fondamentaux des personnes physiques. Si le compromis politique du 8 décembre 2023 paraît confirmer cette avancée, il limite toutefois les interdictions au secteur de l'emploi et de l'éducation. Le Règlement sur l'Intelligence artificielle pourrait donc ne saisir que partiellement le recours à la reconnaissance émotionnelle alors que ses enjeux majeurs, tant sur le plan individuel que collectif, supposent un cadre juridique exigeant et efficient.

Plus généralement, ces enjeux nous invitent à réfléchir à la nécessité d'évaluer si les droits fondamentaux sont suffisants pour garantir le respect de l'autonomie cognitive de la personne et de son intégrité psychique face aux défis posés par les avancées de l'intelligence artificielle, compte tenu en particulier du déploiement massif des agents conversationnels.

On notera à cet égard que des réflexions importantes sur les «neuro-droits» des personnes sont en cours. Au Chili, par exemple, en 2021, le Sénat a voté à l'unanimité en faveur d'un projet de loi visant à modifier la Constitution pour protéger les «neuro-droits». Des discussions similaires sont menées, dans un contexte différent, par la Neurorights Foundation qui examine la nécessité de consacrer de nouveaux droits humains, notamment dans le cadre de l'émergence de nouvelles technologies cognitives. Il s'agit là d'un sujet essentiel qui nécessite une mobilisation d'envergure.

COMITÉ ÉDITORIAL