https://divergences.be/spip.php?article4377

Revue libertaire internationale en ligne

DIVERGENCES

Nous sommes définis par les lignes que nous décidons de traverser ou d'accepter comme frontières

Systèmes complexes

- Aujourd'hui - 2025 - décembre 2025 -

Date de mise en ligne : jeudi 27 novembre 2025

Que signifie mesurer l’intelligence ? | 16 octobre 2025

Origine The Pointmag

Vu du ciel, le désert à l’extérieur de Santa Fe était une masse escarpée jusqu’à l’horizon. La seule chose non escarpée sur des centaines de kilomètres était une ligne droite, une route vers la ville. C’était comme si quelqu’un avait mis du ruban adhésif de peintre sur une page blanche, esquissé des montagnes, des ombres et des masses de la forêt noire, puis soigneusement retiré le ruban.

L’homme à côté de moi dans l’avion m’a demandé si j’habitais à Santa Fe, et j’ai dit non. Il l’a fait. Et connaissait-il l’Institut Santa Fe, où j’avais été invité à un atelier ? Il se contenta de froncer les sourcils.

Peut-être que vous ne le sauriez pas si vous ne le cherchiez pas. L’Institut de Santa Fe (SFI) se trouve sur une route menant à la sortie de la ville, en passant par de douces collines, et finalement dans les montagnes Sangre de Cristo, ainsi nommées parce que leurs pentes brillent de rouge au lever et au coucher du soleil. Des domaines bas et plats de style Pueblo – symboles de l’argent du pétrole du Nouveau-Mexique et du Texan – sont éparpillés sur les collines de piñons et de genévriers comme des dominos lointains. En passant devant, vous pourriez confondre l’institut avec le méga-manoir d’un autre magnat.

SFI est le foyer spirituel de la science de la complexité, une discipline kaléidoscopique qui étudie les « systèmes adaptatifs complexes » – un terme vague pour les systèmes avec de nombreuses parties en interaction, d’où émergent des comportements que nous ne pourrions pas prédire sur la base d’un composant individuel. Le corps humain, les écosystèmes, l’économie mondiale et, de plus en plus, l’intelligence artificielle sont tous présentés comme des exemples de tels systèmes.

En tant que personne qui étudie les impacts sociétaux de l’IA dans une entreprise d’IA, j’ai été invité à un atelier au printemps dernier intitulé « Mesurer l’IA dans le monde » pour aider à « façonner une approche plus scientifique de la mesure des systèmes d’IA ». Le moment était urgent : comme les systèmes d’IA basés sur de grands modèles de

langage sont utilisés par des millions de personnes à des fins tout aussi nombreuses, les chercheurs doivent créer et partager des informations sur leur fonctionnement, ce à quoi ils sont (et ne sont pas) utiles et à quoi pourrait ressembler leur impact sur la société, afin que les gens puissent prendre des décisions éclairées sur la façon d'interagir avec eux.

Bien que les enjeux du déploiement approprié de l'IA semblent de plus en plus élevés, il existe peu d'approches fiables ou établies pour mesurer l'IA sur une dimension d'intérêt donnée. Les LLM sont aussi complexes, sinon plus que les économies : ils émergent de milliards de paramètres interconnectés et sont entraînés sur de vastes ensembles de données par le biais de processus que nous ne comprenons pas entièrement. Comme d'autres systèmes complexes, leurs comportements sont difficiles à prévoir, et pourtant nous avons besoin de moyens robustes pour les évaluer à mesure qu'ils deviennent plus largement utilisés.

En fonction de l'utilisation de l'IA, on peut se demander si elle est aussi familière en japonais qu'en anglais, dans quelle mesure son comportement est cohérent à travers de légères variations d'invites ou si elle peut extraire avec précision des informations pertinentes à partir de documents juridiques. Tous ces éléments nécessitent des approches de mesure plus complexes que l'interrogation du modèle sur des faits facilement vérifiables ; Il n'y a tout simplement pas encore de normes sur ce à quoi cela devrait ressembler.

Prenons l'exemple de l'IA dans les soins de santé : les références existantes de l'IA médicale, comme MedQA, testent les connaissances des manuels scolaires au moyen de questions à choix multiples, mais elles ne capturent pas nécessairement ce qui est nécessaire pour les applications dans le monde réel : des compétences de raisonnement clinique réalistes, une synthèse précise des résultats d'études disparates ou la capacité de décider quand recommander de demander des soins médicaux professionnels au lieu de s'appuyer sur un chatbot. Cet écart de mesure est quelque chose que j'essaie de combler au travail. Mon travail consiste à déterminer quelles dimensions de l'IA sont importantes à comprendre – à la fois pour construire des systèmes d'IA plus bénéfiques et pour s'assurer que la société dispose des informations dont elle a besoin sur ces technologies – et à chercher des moyens de les mesurer : sonder les modèles pour trouver (et corriger) les comportements problématiques, analyser les modèles d'utilisation et de mauvaise utilisation du monde réel, interroger les gens sur la façon dont l'IA affecte leur vie.

La mesure de l'IA est un nouveau domaine, et tout est encore en litige, non seulement la façon dont nous testons, mais aussi ce que nous devrions tester. Tout au long de l'atelier, parmi les crêtes désertiques en forme de croûtes de pain, les participants se sont interrogés sur des questions fondamentales telles que : Comment évaluer si l'IA « raisonne » comme le font les humains ? Est-ce « vraiment intelligent », mais qu'est-ce que cela signifie ? Même si nous ne comprenons pas son fonctionnement interne, pourrions-nous prédire avec précision son impact avant de le déchaîner sur le monde ?

Les préoccupations fondamentales de l'atelier étaient directement liées à mes questions persistantes sur mes propres efforts quotidiens : Comment aborde-t-on la tâche de comprendre les implications d'une technologie entièrement nouvelle ? Les organisateurs ont déclaré que l'utilisation d'approches de mesure issues de domaines existants, à savoir la science de la complexité, pourrait rendre les évaluations de ces nouveaux systèmes plus cohérentes et plus faciles à gérer. J'ai donc réservé un vol, dans l'espoir d'apprendre de domaines qui ont longtemps été aux prises avec des questions désordonnées.

Le premier jour de la conférence, l'autobus de l'hôtel s'est approché de la porte d'entrée de SFI. À l'intérieur, la lumière se déversait à travers les fenêtres du sol au plafond ornées d'équations griffonnées au marqueur lavable argenté. Le soleil rebondissait sur des peintures abstraites lumineuses et illuminait une vitrine remplie de livres de SFI Press, avec des couvertures à la typographie nette et des titres fantaisistes comme The Quark et the Jaguar.

Nous, les participants, nous nous sommes faufilés dans le couloir en tenue d'affaires décontractée, avons mélangé du café noir avec de la crème industrielle et nous nous sommes jetés un coup d'œil sur les badges de l'autre. J'ai reconnu des informaticiens, des spécialistes des sciences sociales, des spécialistes des sciences cognitives, des philosophes ; certains dans l'industrie, la plupart dans le milieu universitaire. Nous nous sommes réunis dans une grande salle de conférence avec des tables circulaires et des chaises en plastique, flanquée de portes-fenêtres à travers lesquelles des mesas déchiquetés étaient visibles au loin.

La séance d'ouverture a révélé un certain nombre de défis auxquels est confrontée la mesure de l'IA. Les gens ne sont pas d'accord sur les définitions de base, ne font pas confiance aux méthodes des autres et ont du mal à transformer un comportement nuancé de l'IA en simples scores bons ou mauvais que beaucoup veulent. Le premier dilemme s'est posé lorsque quelqu'un a soulevé la question (classique) de la façon dont les technologues aiment définir l'« intelligence » – comme, de manière problématique, la capacité d'accomplir des tâches. Un spécialiste des sciences cognitives avec un indice H enviable a soupiré. « Plus vous êtes pragmatique et technique, plus vous êtes conceptuellement superficiel. » Quelques personnes ont ri.

« Ce n'est pas intelligent s'il s'agit simplement d'accomplir des tâches », a convenu un autre professeur. « À moins qu'il ne puisse se déplacer dans le monde de manière fluide, choisir les tâches à accomplir comme un humain le peut, il n'est pas intelligent. » Je me suis demandé pourquoi se déplacer dans le monde avec fluidité et choisir des tâches n'était pas une tâche comme les autres.

Le professeur venait de présenter une conférence disséquant les capacités de résolution de problèmes d'une IA qui jouait à un jeu de société spécifique. Elle a montré comment il mémorisait des modèles plutôt que de développer des règles compressées et abstraites capturant les principes sous-jacents du jeu. Pour elle, cela suggère que l'IA n'a pas la compréhension robuste que les humains développent par l'abstraction, ce qui est crucial pour une véritable intelligence.

Sa revendication m'a semblé rassurante jusqu'à ce que mes pensées commencent à dériver. Qu'est-ce qui est considéré comme une abstraction, et sont-ils vraiment nécessaires à la pensée ? Mon cerveau semble comprendre le monde en piochant de vagues images et des envies dans un tas de vibrations et en les lançant comme des confettis arc-en-ciel. La plupart du temps, il sautille, se débrouillant entre les associations. Cela fonctionne bien jusqu'à ce que j'écrive, et je force ensuite des abstractions, comme des jeans mal ajustés, sur mon brouillage d'impressions.

Notre pause déjeuner a été longue, alors la plupart des participants ont décidé d'aller se promener dans les collines. J'ai marché péniblement dans la poussière, me sentant étourdi, essayant de garder un œil sur les rochers inégaux tout en appréciant la vue. Il n'avait pas l'air d'être si haut, alors j'ai attribué ma sensation d'hébétude à un sommeil limité et à une position assise trop longue. Mais j'ai quand même cherché sur Google l'altitude : 7 500 pieds. Hein ? Nous étions simplement debout sur des collines envahies par la végétation. Je n'arrêtais pas de pointer du doigt le fond de la vallée, qui était juste là, répétant la statistique de 7 500 pieds. Peut-être mes facultés d'estimation m'avaient-elles fait défaut.

Ou peut-être que toute la région était assise sur une mesa géante, une énorme table naturelle. Je me tenais debout, la tête battante, à une hauteur que je ne reconnaissais pas, essayant sans succès de trouver un point de référence qui aurait pu avoir un sens.

Pendant des heures, nous avons discuté de la façon de tester l'intelligence dans l'IA : des expériences plus contrôlées, la création de rubriques pour la mesure de l'IA fondées sur les cadres de sciences cognitives existants, ou l'exploration de la façon dont l'IA est arrivée à son résultat, et pas seulement de sa capacité à cracher la bonne réponse. L'un des problèmes était que les tests auparavant considérés comme nécessitant de l'intelligence, comme

jouer aux échecs ou reconnaître la parole, ont maintenant été rejetés comme ne nécessitant que des heuristiques, des raccourcis, et non une véritable intelligence, maintenant que l'IA les a maîtrisés. Alors, les gens se sont demandé quelle évaluation nous allions mettre en place à l'avance sur laquelle tout le monde serait d'accord et que nous ne changerions pas.

Lors de la séance d'ouverture de la deuxième journée, l'organisateur de l'atelier a fait appel à un éminent universitaire qui n'avait pas pris la parole hier. Alors que nous sirotions un café filtre et piquions nos omelettes, il s'est levé.

« Nous ne pensons jamais à donner des tests de QI à des personnes extraordinairement compétentes », a-t-il déclaré. « Administrer un test de QI à Marie Curie ou à Albert Einstein serait stupide. Je soupçonne qu'Einstein s'en sortirait très mal, probablement surpassé par un jeune garçon de quatorze ans précoce et irritant.

« L'examen par les pairs est également largement critiqué », a-t-il ajouté. « Pour la production experte, il n'existe pas de forme d'évaluation généralement acceptée. » Il s'est assis, nous nous sommes regardés l'un l'autre et la conférence a continué.

Nous avons continué à plaider des stratégies d'évaluation de l'IA, sans jamais répondre à la question suivante : si nous n'avons jamais conçu un test pour l'esprit humain qui capture ce à quoi il aspire, pourquoi croyons-nous que nous pouvons mesurer de manière significative les systèmes d'IA, qui sont beaucoup plus étrangers ?

Quand j'étais à l'université, les Lettres à un jeune poète de Rilke m'ont aidé à surmonter mes angoisses existentielles débilitantes sur ce que je devais faire de ma vie. Rilke m'a assuré que je pouvais essayer de « vivre les questions » qui me tourmentaient, même essayer d'aimer les questions – et peut-être qu'alors je « progressivement, sans m'en rendre compte, je vivrais un jour lointain dans la réponse ».

Mais c'est une chose contre nature à faire. L'esprit humain n'aime pas vivre dans des questions. Un concept en psychologie appelé fermeture cognitive décrit comment notre esprit se précipite pour obtenir des réponses claires et fermes et des résolutions pacifiques à des questions, tout cela pour éviter la douleur de l'ambiguïté. Trouver une relation raisonnable avec la fermeture cognitive est particulièrement nécessaire, et particulièrement difficile, pour les scientifiques. Le travail consiste à trouver des réponses aux questions non résolues, mais le succès dépend de la recherche des bonnes réponses, pas seulement des plus belles.

Je connais un chercheur prolifique en IA qui a été élevé comme un fervent catholique, mais qui a quitté l'Église à la fin de son adolescence. Il travaille sur l'alignement de l'IA, le domaine dédié à s'assurer que les systèmes d'IA agissent conformément aux intentions humaines. Une fois, je lui ai demandé pourquoi il s'était lancé dans la recherche en IA, et il a dit qu'il l'avait fait parce qu'il voulait créer un oracle plus intelligent que les humains, pour lui dire les réponses aux questions métaphysiques qu'il n'était plus capable de trouver en Dieu. « Par exemple, quel est l'ensemble correct de principes moraux ? »

« Vous pensez qu'il y a une bonne réponse à cela ? » J'ai dit. « C'est fou. »

Il cligna des yeux. "Oui, je pense que oui. Peut-être que l'esprit humain ne peut pas le savoir maintenant, mais une superintelligence le pourrait.

Il croit en une vérité supérieure qui dépasse la compréhension humaine mais qui est disponible pour un esprit supérieur (bien que créé par l'homme). Et il veut garder cet esprit en laisse, sous le contrôle humain. Qu'est-ce que

cela signifierait de faire apparaître une vérité que nous ne pouvons pas comprendre ou valider ? Fait-il confiance à l'IA, ou fait-il finalement confiance aux êtres humains ?

La vérité doit être accessible mais transcendante, supérieure mais subordonnée. Pourquoi nos recherches de la vérité fondamentale semblent-elles si souvent contradictoires ?

Dans la deuxième session de l'atelier, « Provocations », un biophysicien théoricien qui a un poste de longue date à SFI a comparé le développement des systèmes d'IA à l'histoire évolutive de la vie sur Terre. En biologie, des composants cellulaires simples évoluent par recombinaison dans des architectures de plus en plus complexes – bactéries, eucaryotes, organes, individus, villes – avec des transitions surprenantes et imprévisibles entre eux. Une architecture donnée définit certaines contraintes sur les comportements possibles pour la vie ; Franchir les barrières d'un régime nécessite donc des changements radicaux dans l'architecture.

« Mais nous n'avons pas de bonnes théories pour comprendre ce qu'il y a de l'autre côté de la barrière », a-t-il ajouté. En biologie évolutive, nous pouvons le voir rétrospectivement, mais nous ne pouvons pas le prédire à l'avance.

Une IA est en effet plus cultivée que conçue, plus comme une plante que comme une machine prototypique. Chaque fois que j'entraîne un modèle d'IA, je suis frappé par la méta-valeur de nos choix, centrés sur la qualité et l'efficacité du processus d'apprentissage (lire : en croissance). Nous arrosons l'IA avec des calculs et des données et regardons d'énormes réseaux de nombres se multiplier encore et encore à la recherche des bonnes architectures internes pour résoudre les problèmes qui lui sont confiés. Nous sommes assis là à espérer que les capacités souhaitées se développent, inévitablement surpris de ce qui émerge.

Essayer de donner un sens à cela d'une manière significative, c'est comme essayer de comprendre la psychologie humaine au niveau cellulaire. Bien que la mesure des propriétés semblables à celles de l'esprit devrait théoriquement être plus facile pour l'IA – les systèmes d'IA peuvent être disséqués et examinés d'une manière que le cerveau humain ne peut pas être – cet avantage est en grande partie théorique. Des milliards de paramètres répartis sur plusieurs couches de calcul numérique créent un réseau trop vaste pour la compréhension humaine. Nous pouvons dessiner un graphique impressionnant des entrailles d'une IA, avec des galaxies de neurones artificiels en action, mais comment allons-nous le comprendre ?

Comme l'esprit humain, l'IA apprend mieux par l'expérience. J'ai appris à jouer au tennis en forant les coups droits et en regardant des pros sur YouTube, pas en lisant des livres de tennis. Une IA apprend également en voyant des exemples, et non en intériorisant directement une logique exprimable. Cette dernière façon de construire l'IA n'a jamais vraiment fonctionné pour nous. Peut-être parce que le type de raisonnement – qu'il soit artificiel ou biologique – adéquat pour résoudre des problèmes complexes doit être si multiforme, si nécessairement constitué par l'expérience, que même s'il est schématisé avec précision, il peut ne pas avoir de sens en tant que fil logique. Il peut s'agir d'un nœud de cheveux monstrueux d'heuristiques et d'impressions, naviguant ordinairement à la vitesse de l'intuition.

Nous ne pouvons garder que quatre ou cinq éléments dans la mémoire de travail à la fois. Nos pensées auraient rampé à dix bits par seconde. Nous ne pouvons probablement pas faire rentrer notre propre raisonnement dans notre tête.

En septembre 2023, un webcomic quotidien populaire intitulé Saturday Morning Breakfast Cereal mettait en scène une femme déterminée faisant irruption dans un immeuble de bureaux. « Toute ma vie, j'ai voulu comprendre ce qu'est la conscience. Maintenant que nous pouvons construire des esprits artificiels, nous allons enfin obtenir une

réponse. Elle poursuit, marchant triomphalement devant une salle remplie de serveurs informatiques : « Plus de dualisme. Plus de mystère. Pas besoin d'agiter les mains et de dire « propriétés émergentes ». Mais : « Oh, nous ne savons pas pourquoi ça marche », dit l'un des ouvriers. Les réseaux neuronaux sont des magiciens.

Il existe un concept en philosophie également appelé fermeture cognitive. Contrairement au concept correspondant en psychologie, il ne décrit pas notre besoin de réponses, mais notre incapacité à y accéder. Le philosophe Colin McGinn a inventé le terme pour indiquer que certaines questions philosophiques peuvent dépasser la compréhension humaine – la question de la conscience, par exemple. La conscience fonctionne comme un moyen de pensée et de perception du monde extérieur, plutôt que comme un objet de représentation lui-même. Un appareil photo ne peut pas photographier son propre intérieur.

Et en effet, les questions croissantes sur la conscience de l'IA ne font que souligner l'inconnaissabilité de la conscience en général. Nous ne pouvons même pas prouver que les autres humains ne sont pas des « zombies philosophiques » manquant d'expérience intérieure, et encore moins de saisir la conscience animale après avoir coexisté avec des animaux pendant des millénaires. Si nous ne pouvons pas confirmer définitivement l'expérience de première main d'un être organique au-delà de nous-mêmes, comment pouvons-nous espérer comprendre la conscience potentielle dans le silicium, un substrat et une forme complètement différents ?

Les chercheurs biomédicaux ne savent pas avec certitude ce que vivent les animaux de laboratoire, mais ils ont, grâce à l'observation, affiné des directives empiriques pour prévenir la détresse : par exemple, la chaleur peut stresser les lapins, et les souris sont plus heureuses lorsqu'elles sont socialisées ensemble. En pratique, nous avançons, accumulant des heuristiques, cherchant comment vivre avec d'autres êtres humains et non humains.

Notre tendance à débattre de termes philosophiques surchargés provient en partie des concepts controversés d'« intelligence artificielle » et de son nouveau frère, « intelligence artificielle générale » (AGI). Certains chercheurs ont déclaré que le nom même de l'IA est le péché originel du domaine, encourageant la pensée anthropomorphique et l'amalgame de capacités disparates sous un même parapluie chargé. Il positionne également imprudemment l'intelligence humaine comme le but ultime.

L'IAG aggrave le problème. Désormais l'objectif déclaré de nombreuses entreprises d'IA, il ajoute un autre mot vague, « général », dans le mélange. L'AGI est le genre de terme qui est juste assez vague et ambitieux pour être utile dans l'industrie de la technologie (« Égalons l'intelligence humaine – sur tout ! »), ce qui en fait également un punching-ball intellectuel pour les universitaires – une personne à l'atelier a déclaré que la poursuite d'un « système à usage général » était une « course folle », la généralité étant quelque chose qui vous éloigne de tout utilisateur particulier. interaction ou contexte.

Lors de l'atelier, nous avions l'impression de tourner en rond et d'éviter ce terme, qui planait comme un nuage d'orage sur nos discussions. « Pouvons-nous simplement évaluer l'artefact tel quel, et non par rapport à l'AGI ? », a demandé quelqu'un. Un autre a évoqué l'aphorisme de Carl Sagan, selon lequel « des affirmations extraordinaires nécessitent des preuves extraordinaires », suggérant que le domaine avait fait d'immenses proclamations trop tôt dans son existence.

Le problème de la définition est profond. Un article récent de la MIT Tech Review a tenté d'aborder le même problème que l'atelier de Santa Fe : comment nous pourrions construire de meilleures évaluations de l'IA. La principale recommandation de l'auteur était de se tourner vers les outils des sciences sociales, où « il est particulièrement important que les mesures commencent par une définition rigoureuse du concept mesuré par le test ». Si nous voulons mesurer à quel point une société est démocratique, par exemple, nous devons d'abord définir la « société démocratique », puis établir des questions pertinentes pour cette définition.

J'ai réfléchi à ce plan en deux étapes. Les définitions semblaient faire carrément partie du problème. Nous pourrions (pourrir) être en mesure de parvenir à une définition consensuelle acceptable de la démocratie, mais l'idée de créer une définition rigoureuse de l'intelligence, ou du raisonnement, est intimidante et pleine de pièges.

J'ai envoyé l'article à une amie qui faisait un doctorat en statistique, et elle m'a répondu par texto : « C'est notoire, les sciences sociales sont terribles pour les évaluations. »

Si nous ne pouvons pas comprendre ce que c'est, comment c'est « penser », si c'est « raisonner » ou « conscient » (et nous avons des doutes incroyables sur notre capacité à poser ces questions aux êtres humains aussi), pouvons-nous au moins essayer de prédire et de circonscrire comment cela pourrait avoir un impact sur notre monde avant de lâcher la chose ?

Les systèmes d'IA contemporains, comme les humains, ne sont pas très bien contraints. Ils apprennent à générer un texte réaliste en lisant de grandes quantités de celui-ci, en absorbant des modèles statistiques implicites qui ne reflètent pas nécessairement les intuitions humaines. Leurs résultats sont flexibles, ouverts et capables de répondre de manière adaptative dans de nombreux contextes différents. C'est en partie la raison pour laquelle le « G » litigieux de l'AGI est qu'ils peuvent généraliser et faire des choses pour lesquelles ils n'ont jamais été spécifiquement formés. Un expert en politiques et en prévisions présent à l'atelier a fait remarquer qu'une partie du défi de la mesure réside dans le fait que nous semblons être entrés dans un « monde de généralité émergente », où la formation sur et pour tout semble en quelque sorte mieux fonctionner que d'essayer d'optimiser pour un ensemble de cas d'utilisation spécifiques et limités.

Évaluer l'impact complet d'un système d'IA avant son déploiement dans le monde réel est donc futile. Des gens comme moi et les autres participants à l'atelier veulent prendre une décision de non-adoption sur ces modèles. Nous les soumettons à une batterie de tests de pré-déploiement, visant à couvrir un maximum de contextes. Bien que de nombreux comportements indésirables puissent être détectés lors de tests internes, même le bac à sable le plus sophistiqué ne peut pas capturer tout ce qui se manifestera après que vous l'aurez envoyé dans le monde pour interagir avec des ordres de grandeur plus humains, dans des ordres de grandeur plus de contextes. Il y a toujours le risque, aussi minime soit-il, qu'un chatbot de service à la clientèle fraîchement déployé, poussé de la bonne façon, vous propose de vous vendre un Chevrolet Tahoe 2024 pour un dollar ou vous suggère de quitter votre femme.

Pour comprendre ce dont quelqu'un est vraiment capable, il faut l'observer alors qu'il sort dans le monde et fait quelque chose de la vie qui lui a été donnée. Nous avons obtenu la véritable mesure de l'impact d'Einstein non pas en le testant sur son QI, mais en le regardant inventer des théories de la physique.

Peut-être devons-nous laisser l'IA se répandre dans le monde pour vraiment la comprendre – une version exacerbée du dilemme de Collingridge : l'idée que l'impact d'une technologie ne peut pas être facilement prédit tant qu'elle n'est pas largement développée et largement utilisée, mais une fois que cela se produit, il devient difficile de la contrôler ou de la modifier.

En pratique, l'IA est souvent mesurée par l'interaction. Les gens jouent avec un chatbot pendant un jour ou deux et décident de lui faire confiance ou non, en s'appuyant sur une expérience qualitative et non sur des mesures quantitatives, un peu comme apprendre à faire confiance à un nouvel ami au fil du temps. Nous ne pouvons pas scruter leur cerveau ou quantifier leur fidélité. À mesure que nos technologies deviennent aussi complexes que nous, nous pouvons considérer un retour au qualitatif, à l'expérientiel et à l'esthétique comme le moyen le plus approprié de donner un sens au monde changeant qui nous entoure.

La tension entre les connaissances quantitatives et qualitatives est aussi vieille que la civilisation. Comme le détaille

Shigehisa Kuriyama dans son livre L'expressivité du corps, les anciens médecins grecs et chinois ont développé des approches radicalement différentes de la conceptualisation et de la lecture des pouls. Les Grecs réduisaient le pouls au rythme, à la fréquence, à la vitesse, à la taille – un battement singulier et dénombrable pour le cœur.

Les praticiens chinois ont distingué de multiples « pouls » qui révélaient la santé de différents organes, les décrivant dans des termes que les Grecs auraient trouvés dérangeants au sens figuré : « rugueux », comme « sable trempé par la pluie », ou « glissant », comme des « perles roulantes ». Les textes grecs ont défini le pouls en grande partie dépourvus d'interprétation ; Les textes chinois ne transmettaient que l'interprétation, seulement l'expérience de la prise de pouls, jamais ce que le pouls « était ». Les Européens qui ont hérité de l'approche grecque ont trouvé que le « maillage dense et enchevêtré de sensations interdépendantes et interpénétrées » que les médecins chinois naviguaient menaçait une science sûre. Pourtant, même leur certitude a parfois vacillé : le médecin français Théophile de Bordeu était d'accord avec les Chinois, disant que le pouls ne pouvait être connu que par le toucher, « par l'expérience et non par le raisonnement, de la même manière que l'on en vient à connaître les couleurs », comme l'écrit Kuriyama.

Lorsque tous les derniers modèles d'IA réussissent tous les tests standardisés que nous avons effectués, nous nous retrouvons à partager des captures d'écran sur les réseaux sociaux pour capturer leurs différences ; rechercher « grande odeur de modèle » sur X et vous trouverez des gens discutant de l'ambiance particulièrement inquantifiable d'un modèle très capable. Les ingénieurs décrivent le travail avec l'IA selon qu'ils ont l'impression de collaborer avec un collègue senior ou junior. Lorsque les mesures échouent, nous nous tournons vers une métaphore familière : les êtres humains.

Cormac McCarthy a été membre de l'Institut de Santa Fe pendant des décennies, un auteur parmi les mathématiciens et les physiciens. Il avait l'habitude de travailler sur sa machine à écrire Olivetti sur un grand bureau en bois dur dans la bibliothèque SFI. Dans ses derniers romans, The Passenger et Stella Maris, les personnages utilisent des preuves mathématiques comme lentilles sur la nature de la réalité, la mécanique quantique et la conscience. Le protagoniste de Stella Maris est un jeune génie mathématique brillant et psychologiquement désordonné, pour qui les mathématiques apparaissent à la fois comme un salut et une malédiction. Il lui offre une précision dans un monde d'ambiguïté, mais son exactitude même souligne combien de choses restent inconnaissables. Elle s'enfonce dans une spirale.

Pour moi, le grand attrait de la science de la complexité est sa volonté de tenir compte du mystère émergent dans le monde, d'essayer de l'entourer de ses bras, de l'affronter courageusement. Pourtant, debout dans la bibliothèque de la SFI en feuilletant les pages reliées par des cas des documents fondamentaux en science de la complexité, laissant mon esprit glisser sur les formules et les diagrammes, je suis frappé par la façon dont la terminologie de la science de la complexité est principalement mathématique, cherchant à réduire la complexité en équations obéissantes qui confinent l'enquête aux plus hauts niveaux d'abstraction. Je me demande si la fiction n'a pas offert à McCarthy ce que les formules n'ont pas pu offrir : un langage pour les aspects ineffables de la complexité, lorsque les systèmes que nous étudions confondent nos instruments savants. Sa prose directe, évocatrice et quasi psychédélique coupe un couteau dans ces facettes de notre expérience que la science a du mal à expliquer.

Au cours des deux jours à Santa Fe, nous avons établi certains problèmes et certains principes : il est difficile d'évaluer les outils à usage général utilisés dans d'innombrables contextes, et nous avons besoin d'évaluations spécifiques au contexte mais systématiques. La prédiction de nouveaux systèmes dans de nouveaux contextes est difficile ; Nous avons du mal à mesurer même ce qui semble simple. Mais que devrions-nous faire à ce sujet ?

Au cours des dernières sessions de l'atelier, nous nous sommes divisés en petits groupes pour travailler sur des sujets spécifiques, puis nous nous sommes réunis pour les présenter. Une pensée me trottait dans la tête : et si nos systèmes de mesure étaient aussi des systèmes complexes ? Qu'en est-il de quelque chose qui peut s'accrocher à

l’arrière du problème et s’adapter à sa taille ? Mon groupe a préparé quelques notes sur l’idée d’évaluations post-déploiement pour comprendre les impacts sociétaux de l’IA – des systèmes de signalement des incidents qui répertorient les problèmes au fur et à mesure qu’ils se produisent, des initiatives de science citoyenne, des façons dont notre système de mesure peut, au lieu de fixer un cadre spécifique et d’insister sur des réponses a priori, être adaptatif et co-évoluer avec notre société technologique.

Cela ne répondrait pas aux questions fondamentales sur ce qu’est un système d’IA ou sur son fonctionnement, mais cela pourrait nous aider à comprendre ce que nous en faisons, en sortant de l’environnement de laboratoire propre et en travaillant dans le cadre de nos contraintes, plutôt que de les désespérer. Les technologies mêmes que nous inventons pour étendre nos connaissances, les puissants projecteurs que nous visons dans l’obscurité, révèlent également les contours distincts de tout ce que nous ne pouvons toujours pas savoir.

