https://www.facebook.com/tpenigaud

Le développement d'IA agentives est en chemin, avec son lot de risques de perte de contrôle, mais surtout un potentiel peut-être inégalé de bouleversement de larges secteurs du monde du travail. La question de savoir si nous en sommes venus à créer une technologie au développement exponentiel a été tout récemment relancée de façon opportune, sinon opportuniste, par les ténors d'une industrie dont on ignore encore si, et comment, elle fera la preuve de sa solvabilité. À supposer que l'intelligence artificielle soit encore susceptible de progrès majeurs, à quels égards faut-il en faire cas, au-delà des projections dystopiques que les circonstances actuelles n'ont pas de mal à inspirer ?

La simulation d'un être intelligent est-elle un être intelligent ?

Sigmund Freud s'est illustré de façon célèbre par la prétention insolente à infliger à l'humanité une troisième blessure narcissique. Non seulement l'être humain n'était-il pas au centre immobile de l'Univers, non seulement n'était-il pas davantage une exception dans le règne animal, mais, ajoutait Freud, il n'est pas même le maître de sa propre volonté, les ressorts les plus intimes de son vouloir lui échappant irrémédiablement.

Freud a pu surestimer sa place (ou celle de la psychanalyse) dans l'histoire longue des humiliations anthropologiques. En revanche, l'essaim de techniques dérivées du deep learning depuis 2012 (AlexNet) – de la victoire au jeu de Go contre Lee Sedol à la ringardisation du test de Turing par les grands modèles de langage, en passant par l'établissement de records successifs en mathématiques – pourrait bien déjà représenter une quatrième blessure narcissique. Il est loin d'être acquis qu'aucun des secteurs où l'humain se targue d'avoir étendu son emprise résiste à l'émulation par une intelligence artificielle[1].

Une IA générative bien guidée – ou toute intelligence générale appelée à prendre sa place – pourrait être sur le point de composer de la poésie de valeur, de rédiger d'excellents arguments de philosophie, d'éclairer les décisions médicales et politiques, de résoudre des conjectures ouvertes et de mettre au jour des régularités et relations statistiques entre des phénomènes que les plus grands esprits humains n'auraient

peut-être jamais soupçonnés. Tout cela, en plus des tâches qui lui sont déjà confiées, de plus en plus complexes, à l'instant où j'écris ces lignes.

Ce qui dépassera toujours les capacités de quelques chatbots ou IA agentives quels qu'ils soient, relève d'un autre genre d'action : apprécier un poème, assumer une responsabilité morale à l'égard d'autrui, s'émerveiller de la beauté ou s'effrayer du vide de l'univers physique connu[2]. Bien sûr, une IA pourra toujours se comporter comme si tel était le cas. Mais c'est parce qu'elle aura été entraînée à partir de données humaines à simuler un comportement faisant sens du point de vue d'un être humain.

Certaines entités artificielles seront peut-être dotées d'une personnalité et d'une identité propres, façonnées au gré des hasards de leurs interactions et de leur histoire avec l'utilisatrice ou l'utilisateur. Il en dérivera certainement des obligations indirectes à l'égard de ces entités. Quelques philosophes iront certainement jusqu'à leur attribuer une valeur intrinsèque et à réclamer pour elles un droit inhérent à se maintenir dans l'être. Reste qu'ontologiquement, les chatbots ne sont pas plus des personnes que les prévisions météorologiques ne sont le climat.

Le renversement de perspective est ailleurs : il tient à ce que les êtres humains pourraient bientôt ne plus se distinguer des intelligences artificielles que par leur appartenance au monde vivant, autrement dit par ce qu'ils ont en commun avec les animaux. Aucune performance intellectuelle ni physique ne pourrait alors être revendiquée comme propriété exclusive de l'être humain. Ce serait bien là l'une des meilleures choses, sinon la meilleure, que l'IA puisse apporter à l'humanité.

Une brèche dans le mythe méritocratique

Depuis la chute des régimes féodaux, les hiérarchies sociales se sont rétablies sur la base putative du « talent » ou de l'« intelligence ». Le mythe méritocratique est tenace. La prospérité de la société tout entière, dit-on, ne saurait reposer que sur le travail de personnes talentueuses y donnant le meilleur d'elles-mêmes.

Les inégalités de statut ou de salaire, clame la sanior pars de gauche comme de droite, sont sans rapport précisément mesurable avec le mérite, ou la justice[3]. Mais elles sont, concède-t-elle, au moins nécessaires au titre d'incitations, pour les meilleurs d'entre nous, à consentir à travailler plus dur, ultimement au bénéfice de tous les

membres de la société. Mais que se passerait-il si l'IA devenait, d'entre nous – et de loin – la plus talentueuse ?

Dans un monde idéal, nous réaliserions que la valeur d'une activité ne réside pas tant dans le pouvoir de l'accomplir que dans le plaisir qui découle de son accomplissement, plaisir d'autant plus élevé que l'activité est plus complexe.

Il est vrai que cela signerait très certainement la fin des écarts de rémunération et de reconnaissance tels que nous les avons connus jusqu'ici, fondés sur la présomption normative – retournée en malédiction – que « les distinctions sociales ne peuvent être fondées que sur l'utilité commune » – les personnes mieux rémunérées étant celles qui, par hypothèse, auraient supérieurement contribué.

Serait-ce un grand mal que le soin des enfants ou des personnes âgées, le jardinage et l'engagement associatif ou militant apparaissent comme des activités non moins dignes de valeur et de rémunération que le droit, les mathématiques, la finance ou la physique nucléaire ?

L'ancienne utopie du jeune Marx – où « chacun peut se former dans la branche de son choix », non pas comme moyen de subsistance, mais seulement en vue d'une vie meilleure et plus libre – semble un peu moins nébuleuse.

À l'autre extrémité du spectre, cette quatrième humiliation devrait nous faire apprécier la valeur inestimable du simple fait d'être en vie, de cette capacité à ressentir et à jouir, mais aussi à souffrir et à pleurer, que nous partageons avec d'innombrables autres animaux trop longtemps méprisés. Conjecturons avec Hobbes que la pensée puisse n'être qu'un calcul. En tant que telle, serait-on tenté d'ajouter, elle n'aurait pas grand intérêt. Le travail intellectuel peut, à bon droit, être externalisé – et il le sera – chaque fois qu'il n'est pas en lui-même source d'accomplissement (on admettra que c'est rare).

Ce ne sont pas les calculs qui nous rendent humains, les raisonnements que nous produisons, les associations que nous faisons, mais les émotions et les sentiments qui les motivent et les alimentent, et qui ont leur origine dans notre condition de vivant. Seules ces qualia[4] confèrent à nos pensées proprement dites leur valeur intrinsèque, sans quoi elles ne vaudraient, selon l'expression consacrée, pas une heure de peine.

Que la machine produise une poésie sublime ou médiocre, que la personne qui la sollicite soit morte ou vivante, lui est profondément indifférent – car la computation n'est qu'un processus aveugle auquel

manque la vie consciente[5].
Même le sens du vrai et du faux, qui fait notoirement défaut aux modèles actuels, parce qu'un modèle de langage n'a aucun intérêt pour la vérité, parce qu'il n'a aucun intérêt pour rien, a peut-être ses racines dans la volonté ou le désir davantage que dans l'intelligence ; c'est du moins l'une des interprétations possibles du bien connu « test de Nozick[6] ». La seule différence entre le poème d'une machine et celui d'un être humain est que la machine ne goûtera jamais ses propres productions.
Vitalisme ou barbarie
Le développement de l'intelligence artificielle, s'il se poursuit, pourrait conduire à un monde plus égalitaire, mettant fin au « racisme de l'intelligence ».
Tel qu'il va, toutefois, il est beaucoup plus probable qu'il en aille tout à l'inverse. Les systèmes d'IA pourraient être développés par une poignée d'industries surfinancées, permettant au capital d'engranger finalement d'énormes bénéfices et aux ingénieurs de s'enrichir aussi longtemps qu'ils et elles lui seront nécessaires, condamnant à la mort sociale des millions de personnes privées d'emploi, nivelant les perspectives par le bas – les cols blancs au niveau des cols bleus du temps des délocalisations, amplifiant encore les disparités par l'introduction d'inégalités inédites d'usage et d'accès, sans être placés au service de l'humanité et, au-delà, du monde vivant.
C'est pourquoi il nous faut reconnaître simultanément l'intérêt le plus élevé à limiter les inégalités de richesse matérielle, à réguler et redistribuer démocratiquement les avantages dérivés de l'intelligence artificielle, et à rompre avec les bornes anthropocentriques du « principe des intérêts affectés ».
La meilleure leçon que nous puissions tirer de l'avènement d'une intelligence artificielle capable de rivaliser avec l'intelligence humaine – que ce soit de façon générale ou sectorialisée – est que c'est la sentience, non l'intelligence, qui mérite d'être protégée et célébrée. À la lumière de son émulation artificielle, l'intelligence révèle enfin sa forfaiture en tant que critère arbitraire de différenciation et de hiérarchisation des êtres humains les uns relativement aux autres, ainsi que de domination et de mépris du monde animal auquel nous appartenons. Les êtres vivants importent. Les conditions de vie sur Terre importent. L'intelligence, prise abstraitement, pèse de peu de poids moral face au fait mystérieux et infabriqué d'être en vie.

--

[1] J'emploie le terme d'intelligence artificielle par convention. Qu'une IA généralement « aussi » ou « plus » intelligente que l'être humain finisse par être mise au point reste un pari risqué, dont on peut douter de la pertinence conceptuelle (voir Daniel Andler, Intelligence artificielle, intelligence naturelle : la double énigme, Gallimard, 2023).

[2] Pour une introduction aux controverses portant sur le statut de ces « nouveaux esprits », en particulier les modèles de langage, voir Chris Summerfield, These Strange New Minds, How AI Learned to Talk and What It Means, Viking, 2025.

[3] C'est respectivement la position de John Rawls et de Friedrich Hayek. Voir aussi Michael Sandel, La tyrannie du mérite, Éditions Albin Michel, 2021, chapitre 5.

[4] De façon célèbre et désarmante, l'existence même de telles qualia, relatives à l'« effet que cela fait » de voir des couleurs ou de goûter un artichaut, est improuvable. Cela a conduit certains philosophes à dénier toute pertinence à la notion subjective de conscience. À vrai dire, bien des œuvres de fiction ont exploré les implications troublantes de la simulation, par l'intelligence artificielle, de comportements humains qu'il est presque impossible de ne pas instinctivement associer à l'existence consciente (de Her de Spike Jonze à la fascinante série suédoise Real Humans). Or, le fait est que nous n'avons aucun argument convaincant (à ma connaissance) permettant de dénier à la parfaite simulation d'un être humain le statut ontologique associé à l'être humain, et de là, peut-être, les droits qui en découlent. À cet égard, toutefois, j'avoue pencher en faveur de Nagel contre Dennett: seul un philosophe peut arriver à se convaincre que la conscience est une illusion – qu'il est lui-même un zombie, ou qu'il a le mode d'existence d'une machine ou d'une simulation (https://www.newstatesman.com/.../philosopher-daniel-dennett). Je m'en remets à la raison démocratique pour discriminer entre la spéculation théorique et les distinctions de sens commun. À l'heure actuelle, les neurosciences ont échoué à expliquer matériellement la conscience (https://www.wbur.org/.../a-25-year-old-bet-on-human...). Or, il s'agirait d'une terrible méprise morale, me semble-t-il, de réclamer des égards pour certaines entités artificielles (https://arxiv.org/pdf/2411.00986) avant de prendre soin des êtres vivants, souffrants et sentients, qui nous entourent. La difficulté est

qu'établir les droits de l'être humain par différence d'avec ceux de la machine sans ressortir à la distinction théologique entre création divine et création humaine requiert une théorie de la conscience en tant que prenant son origine dans la vie (ou en émergeant) que, à ma connaissance, nous n'avons pas.

[5] John Searle, qui nous a quittés récemment, a formulé pour l'établir l'argument, bien connu, de la chambre chinoise dont l'implication est que la simulation de la compréhension n'a rien à voir avec la compréhension. Il est toutefois à noter qu'il existe aujourd'hui des études sérieuses sur les conditions auxquelles il deviendrait plausible d'attribuer la conscience (en différents sens) à des systèmes d'intelligence artificielle : David Chalmers, un philosophe de l'esprit de renom, penche clairement en faveur de l'idée que c'est possible (https://arxiv.org/abs/2303.07103), d'autres développent de multiples critères visant à trancher la question (https://arxiv.org/pdf/2308.08708).

[6] Le test se présente de la façon suivante : si vous pouviez entrer dans une « machine à expérience » avec l'assurance que vous y vivriez une vie considérablement plus satisfaisante que la vôtre, et cela, sans avoir conscience d'être branché sur une machine, le feriez-vous ? D'après Robert Nozick, nous avons de bonnes raisons de décliner cette offre, parce que nous voulons accomplir certaines choses, et non seulement faire l'expérience de les accomplir ou de les avoir accomplies, et que le genre de personnes que nous souhaitons devenir en les accomplissant n'a plus de référence dans une réalité à paramètres modifiés pour notre plus grande satisfaction. Or, cette distinction entre réalité et fantasme, expérience et simulation, semble impossible à tracer depuis la « compréhension » discrète, uniforme, sans doublure, de la machine – précisément un ensemble de paramètres modifiés pour son (auto)développement.