Traduction par l'ia deepseek de l'article arxiv 2603.26524 de Tanya Klowden et Terence Tao, 7 août 2026

1. Résumé

L'intelligence artificielle (ia) est le nom communément donné à un large éventail d'outils informatiques conçus pour effectuer des tâches cognitives de plus en plus complexes, y compris beaucoup qui relevaient autrefois du seul domaine humain. Alors que ces outils deviennent d'une sophistication et d'une omniprésence exponentielles, les justifications de leur développement et de leur intégration rapides dans la société sont de plus en plus remises en question, en particulier lorsqu'ils consomment des ressources limitées et posent des risques existentiels pour les moyens de subsistance des personnes qualifiées qu'ils semblent remplacer. Dans cet article, nous considérons l'impact en évolution rapide de l'ia sur les questions traditionnelles de la philosophie, en mettant l'accent sur son application en mathématiques et sur les conséquences concrètes plus larges de son usage généralisé. Nous affirmons que l'intelligence artificielle est une évolution naturelle des outils humains développés au cours de l'histoire pour faciliter la création, l'organisation et la diffusion des idées, et nous soutenons qu'il est primordial que le développement et l'application de l'ia restent fondamentalement centrés sur l'humain. En vue d'innover des solutions répondant aux besoins humains, d'améliorer la qualité de vie humaine et d'étendre la capacité de la pensée et de la compréhension humaines, nous proposons une voie pour intégrer l'ia dans nos domaines les plus exigeants et les plus rigoureux sur le plan intellectuel, au bénéfice de toute l'humanité.

1. Introduction

Il est révélateur de voir à quelle vitesse les technologies de l'intelligence artificielle (ia) ont été déployées dans tous les recoins de la vie numérique : alors qu'ils rédigeaient cet article à l'aide d'outils standards, les auteurs ont vu pas moins de trois agents numériques différents s'immiscer dans le récit sans y être invités [1]. L'humanité se tient au seuil d'une révolution industrielle numérique, qui se déroule à des vitesses sans précédent. Dans les sciences physiques, les progrès de l'ia ont mené à des recherches primées par le prix Nobel [1] ; tandis que dans les sciences humaines, on craint que les capacités de génération de texte de l'ia moderne ne signent la mort de la discipline [2]. Alors que les traducteurs linguistiques ont ouvert grand les portes aux échanges culturels et à la coopération internationale, un flot de deepfakes et de contenu de faible qualité (slop) a suivi, déferlant dans nos espaces numériques. L'ia est rapidement passée du statut de nouveauté à celui de ressource vitale, pour devenir (dans certains cas) une menace existentielle immédiate [3].

1.1. Notre définition de l'intelligence artificielle.

Pour les besoins de cet article, l'ia désigne le large spectre d'outils informatiques conçus pour effectuer des tâches cognitives de plus en plus complexes, y compris beaucoup qui relevaient autrefois

1. Toutes ces "contributions" de l'ia ont été immédiatement supprimées du texte.

du seul domaine humain. Les outils d'ia sont extrêmement divers, allant des technologies d'apprentissage automatique (machine learning, ML) pilotées par les données d'aujourd'hui (telles que les grands modèles de langage (llm) capables de traiter des textes complexes, ou les modèles de diffusion capables de générer des images et d'autres médias), jusqu'à l'ia plus traditionnelle dite "à l'ancienne" (gofai, pour good old-fashioned ai) (comme les démonstrateurs automatiques de théorèmes ou les moteurs d'échecs), qui peuvent résoudre des problèmes dans des domaines restreints en appliquant des règles mathématiques précises.

1.2. Objectif de cet article.

Les discussions sur ce que ces outils peuvent ou ne peuvent pas faire ne manquent pas ; mais on discute comparativement moins des raisons pour lesquelles ces outils sont développés et déployés si rapidement, ni de leur impact sur les milliards de vies qui interagissent avec eux pour la recherche et l'éducation, le travail, les loisirs, et même le repos [4]. Les auteurs de cet article viennent de domaines académiques souvent perçus comme opposés : les mathématiques et l'étude de l'art. Mais nous avons tous deux trouvé bénéfique d'incorporer plusieurs outils d'ia dans nos domaines de recherche disparates au quotidien, et nous avons découvert un nombre surprenant de points communs concernant les questions philosophiques très complexes, mais universelles, que pose l'utilisation réelle de l'ia. En prenant les mathématiques comme modèle, nous examinerons les avantages, les risques, l'éthique et les résultats de l'incorporation de l'ia dans les flux de travail courants, puis nous étendrons ces observations à une utilisation plus large dans le monde réel. Malgré les risques que présentent ces technologies nouvelles, et pas nécessairement moralement neutres, nous soutenons une double thèse : les outils d'ia devraient être développés, mis en œuvre et appliqués à la fois en mathématiques et dans d'autres domaines ; ils ont le potentiel d'augmenter radicalement nos capacités humaines naturelles et ils sont capables d'étendre ce qui est possible au-delà de ce que nous, humains, pourrions faire individuellement ou dans les limites de notre propre capacité collective. En nous appuyant sur nos propres expériences avec ces outils, nous examinons en particulier l'interface humain/ia et proposons des suggestions sur l'évolution de ces technologies vers des modalités offrant plus de bénéfices que de préjudices à l'humanité, et valorisant les contributions uniques de la pensée et de l'action humaines en concert avec les nouvelles modalités que promet le développement futur de l'ia.

1.3. Le pacte faustien.

Les incitations de la concurrence marchande ont alimenté un rythme effréné de développement des technologies de l'ia et ont fasciné des industries entières par des visions de flux de travail radicalement accélérés et d'économies de coûts. Le "dilemme du prisonnier" d'une telle concurrence a poussé de nombreuses personnes et organisations à adopter ces outils de manière expérimentale aussi rapidement que possible, au détriment d'une évaluation plus délibérée des coûts et avantages économiques, sociaux ou moraux d'une telle adoption - ou, plus fondamentalement, des raisons pour lesquelles nous devrions développer de telles technologies en premier lieu. En conséquence, nous avons collectivement déjà conclu un pacte *de facto* de type "faustien" avec ces technologies, leur donnant un accès croissant à nos données, nos flux de travail cognitifs et nos processus de décision, en échange de la promesse de pouvoir accomplir une gamme plus large de tâches avec une

efficacité accrue et un moindre effort fastidieux.

En théorie, la technologie est moralement neutre ; elle peut servir des cas d'usage positifs comme négatifs. Mais par cet autonomisation, elle exacerbe les dilemmes moraux existants et en crée de nouveaux. Par exemple, les recherches médicales horribles menées sur des prisonniers pendant la Seconde Guerre mondiale, qui ont fourni des données vitales sur les limites de l'endurance humaine, ont soulevé des questions difficiles quant à l'éthique de l'utilisation de ces données pour développer de nouvelles avancées médicales [5]. Sans être aussi macabre, la provenance douteuse des données et de la propriété intellectuelle utilisées pour entraîner la génération actuelle d'outils d'ia soulève probablement des questions similaires aujourd'hui [6].

Lorsqu'une technologie se développe assez lentement, il est possible d'avoir les conversations et débats philosophiques nécessaires à son sujet avant qu'elle ne soit largement déployée ; la recherche sur les cellules souches en est un exemple notable. Cependant, les technologies modernes de l'ia sont déjà largement déployées, sans moyen pratique de "remettre le génie dans la bouteille" ; ironiquement, une réglementation stricte imposée à ce stade fermerait de manière disproportionnée les cas d'usage les plus positifs de l'ia, comme l'accélération de la recherche scientifique, sans éliminer les utilisations les plus gaspillantes ou malveillantes de la technologie. De manière pragmatique, le débat sur l'ia s'est désormais orienté vers la gestion de la coexistence avec ces technologies : évaluer les coûts et les avantages de l'ia (à la fois dans les disciplines académiques et dans la société en général), et identifier les meilleures pratiques et cadres pour utiliser l'ia de la manière la plus positive possible, tout en décourageant simultanément les (nombreuses) façons dont ces outils peuvent être mal utilisés pour dégrader la fiabilité et la valeur de nos réalisations cognitives.

2. Parallèles historiques : cette fois-ci est-elle différente ?

2.1. Intégration passée des technologies d'automatisation.

L'automatisation n'est bien sûr pas un phénomène nouveau. De nombreuses technologies passées ont également permis d'automatiser des tâches auparavant assignées aux humains, éliminant ou réduisant considérablement le besoin de certains types d'emplois humains, tout en créant ou augmentant le besoin d'autres ; dans certains cas. Au sein de la communauté scientifique, par exemple, des "transitions de phase" se sont produites, où l'on est passé largement et rapidement à de nouveaux outils (comme Internet, l'utilisation des ordinateurs pour le calcul scientifique, ou même de modestes langages de composition comme LaTeX) en raison de leurs avantages évidents. Mais ces technologies passées ont surtout affecté des aspects secondaires de la profession, comme la communication et la diffusion des résultats plutôt que la création de ces résultats. Et, bien que les tâches automatisées par ces outils aient exigé une formation spécialisée et une expertise pour être réalisées, elles ne requéraient généralement pas une compréhension de la dimension plus créative.

2.2. L'ia moderne.

Mais l'ia moderne peut automatiser de grandes parties du processus créatif lui-même, permettant la production de masse de produits intellectuels, tels que des œuvres d'art, des preuves mathématiques ou des théories scientifiques ou philosophiques, avec une supervision humaine bien moindre qu'auparavant [2]. Cela a créé un découplage sans précédent entre la forme extérieure de ces produits et les valeurs et processus de pensée utilisés pour les créer. Un modèle de diffusion peut désormais créer un paysage esthétiquement agréable, par exemple, qui n'a été directement inspiré par aucun lieu particulier du monde physique, bien que d'innombrables images de paysages réels (ainsi que de nombreuses images sans rapport avec les paysages) aient certainement été utilisées pour entraîner les sorties de ce modèle ; la réponse esthétique à l'image devient ainsi découplée des sources originales de ces esthétiques.

Ce n'est pas un territoire philosophique nouveau. L'expérience de pensée de la "chambre chinoise" de Searle [8], concernant la question de savoir si un dispositif mécanique programmé pour converser en chinois comprend véritablement la langue, date de 1980. "L'effet ia" a également été reconnu à cette époque ; par exemple, la capacité à bien jouer aux échecs était considérée comme une bonne mesure de l'intelligence jusqu'à l'avènement des moteurs d'échecs qui pouvaient surpasser "mécaniquement" les grands maîtres par l'exploration systématique des arbres de jeu, moment à partir duquel le "test des échecs" pour l'intelligence a été largement abandonné. Le célèbre "test de Turing" - savoir si une ia pouvait converser d'une manière impossible à distinguer des humains - a été effectivement réussi par les llm modernes (voir, par exemple, [9]), perdant ainsi son ancien statut de "référence absolue" pour l'intelligence artificielle. Pour une discussion plus récente, voir [10].

Pour l'instant, nous pouvons encore pointer des marqueurs de compréhension "fondamentale", comme la capacité (ou l'absence de capacité) à expliquer et défendre de manière cohérente les processus créatifs qui ont mené à une nouvelle œuvre d'art, une preuve mathématique ou un autre produit intellectuel, comme test encore viable pour distinguer le contenu humain de celui généré par l'ia. Mais si les générations futures d'ia parvenaient également à réussir de manière convaincante de tels tests, devrions-nous une fois de plus déplacer les limites de ce que sont réellement l'intelligence, la compréhension et la créativité ? Les définitions, les valeurs et les objectifs de disciplines comme les mathématiques et les sciences humaines devraient-ils être réévalués ? Et quel statut devrions-nous accorder à ces outils d'ia de plus en plus sophistiqués - seront-ils des assistants, des co-auteurs, ou même des créateurs indépendants à part entière ? Et si oui, comment devrions-nous traiter le contenu qu'ils produisent, et les processus intellectuels qui ont mené à ce contenu ?

3. Les mathématiques comme bac à sable pour l'utilisation de l'ia

Ces questions philosophiques plus larges sur l'ia sont extrêmement complexes et multidimensionnelles, et nous ne prétendons bien sûr pas avoir de résolutions définitives à aucune d'entre elles ; et

2. Les outils actuels nécessitent généralement encore l'intervention d'un humain pour générer une consigne initiale que l'ia devra suivre, mais ce processus peut désormais être largement automatisé.

la vitesse du changement dans ce domaine est telle que toute proclamation que nous faisons risque d'être dépassée par de nouvelles avancées technologiques frappantes. Cependant, nous pouvons offrir quelques perspectives du monde des mathématiques, à la fois dans le domaine du raisonnement mathématique pur et dans l'application émergente de l'analyse mathématique moderne dans les sciences humaines. Nous considérons les mathématiques comme un "bac à sable" approprié pour explorer des questions larges telles que l'impact de l'ia à travers les sciences (et la société dans son ensemble), car elles disposent d'une base plus ancienne et plus avancée, et sont par nature bien adaptées pour explorer une variété de scénarios abstraits hypothétiques contrefactuels par rapport à la réalité. Nous espérons que les leçons tirées de l'intégration (ou non) de l'ia dans les mathématiques pourront offrir des perspectives plus larges sur la manière dont l'ia interagira avec les sciences et la société en général.

Les modèles d'ia de pointe peuvent désormais résoudre des problèmes mathématiques de plus en plus complexes, avec des preuves vérifiables de manière indépendante, sans reproduire directement les pratiques de résolution de problèmes des mathématiciens humains (comme tester des cas particuliers, puis généraliser à partir de ces exemples), bien que leurs données d'entraînement incluent des preuves générées de manière traditionnelle ; et les mathématiciens seront donc de plus en plus confrontés à des situations où la capacité à démontrer des théorèmes est découplée des processus de raisonnement nécessaires pour découvrir et comprendre ces preuves. Cela contribue à une tendance existante à la décentralisation en mathématiques modernes ; dans un monde où les mathématiques avancées sont nécessaires dans un éventail extrêmement large d'applications, l'ère "Bourbaki" [11] où une autorité centrale prescrivait la pratique orthodoxe des mathématiques est déjà révolue depuis des décennies [3].

4. L'ia et la nature de la vérité mathématique

En l'état actuel de la technologie, les outils d'ia les plus sophistiqués présentent encore des faiblesses significatives et souvent bizarres ; ils peuvent atteindre des performances remarquables et suprahumaines dans certaines tâches, tout en montrant simultanément des niveaux souvent risibles d'incompréhension fondamentale et d'erreur dans d'autres. Les mathématiques ne font pas exception à ce phénomène. Les mathématiques générées par l'ia peuvent sembler superficiellement impeccables - ce qui est attendu, car ces modèles sont conçus pour produire des résultats aussi proches visuellement que possible des preuves humaines correctes - tout en commettant des erreurs fondamentales (par exemple, affirmant que tous les nombres impairs sont premiers) qui auraient été éliminées chez un mathématicien humain dès les premiers stades de sa formation, et peuvent souvent rendre l'argument résultant dangereusement absurde. En même temps, cette approche "descendante" consistant à se concentrer principalement sur la production de résultats d'aspect satisfaisant plutôt que sur les processus cognitifs fondamentaux traditionnellement nécessaires pour créer ces résultats peut être étonnamment efficace ; la même ia qui commet régulièrement des erreurs mathématiques de base peut aussi arriver mystérieusement à la bonne réponse à un problème mathématique complexe avec une précision supérieure à celle des experts humains, ou même fournir

3. On pourrait certes arguer que les projets en cours visant à créer de vastes bibliothèques unifiées de mathématiques formelles, comme le projet `mathlib` de `lean`, pourraient constituer un successeur moderne aux efforts du groupe Bourbaki.

une preuve étrange mais techniquement correcte que la réponse est valide.

Des efforts considérables sont maintenant déployés pour réduire ou éliminer ces faiblesses de l'ia autant que possible ; souvent non pas en renforçant directement la "compréhension" innée de l'ia pour une tâche intellectuelle donnée, mais plutôt en plaçant ces outils d'ia dans un environnement rigoureux de tests, d'entraînement et de vérification indépendants pour réduire l'incidence numérique des erreurs. La capacité à résoudre des conjectures mathématiques profondes reste actuellement hors de portée d'une ia complètement autonome, mais il est très plausible que dans un avenir proche, de tels outils d'ia puissent grandement assister les mathématiciens humains dans de telles entreprises, même si nous hésiterions encore à décrire une telle assistance comme l'expression d'une pensée véritablement intelligente. Néanmoins, le fait que des approches aussi mécanistes et sujettes à l'erreur pour une discipline aussi intellectuelle que les mathématiques puissent (ou vont bientôt) générer autant de marqueurs traditionnels de qualité dans la discipline indique que nous devons réévaluer nos modèles de ce qu'est réellement l'intelligence ou la créativité, et comment elles doivent être mesurées.

4.1. Mathématiques et standards de preuve.

Les mathématiques [4] ont une longue tradition de standard objectif de preuve, commençant avec Euclide et raffiné avec l'avènement de fondements stables et (empiriquement) sûrs des mathématiques au début du vingtième siècle. Il a été noté (voir, par exemple, [13]) que l'acceptation quasi-universelle de ces fondements a donné aux mathématiques modernes la capacité rare et précieuse de parvenir à un consensus sur la validité d'un argument ou d'une assertion donnée dans le domaine, puisque (en principe) on pourrait insister pour que ces arguments soient exposés avec un tel niveau de détail que chaque étape individuelle puisse être vérifiée comme une application correcte des axiomes standards et des règles d'inférence logique des mathématiques. Un exemple typique a été l'affirmation de Nelson [14] en 2011 selon laquelle les axiomes de Peano étaient logiquement incohérents ; c'était une affirmation très éloignée du courant dominant mathématique, et pourtant il a été possible de résoudre la question en signalant un défaut subtil dans l'argument que Nelson a volontiers accepté, retirant ainsi l'affirmation.

Cependant, dans la pratique, les arguments des mathématiciens humains sont loin de l'idéal de la preuve parfaitement rigoureuse ; les erreurs mineures et majeures dans la littérature sont courantes, certaines étant corrigées par des errata formels ou des révisions, et d'autres étant négligées ou transmises informellement comme faisant partie du "folklore" du sous-domaine. Les arguments qui sont heuristiquement plausibles sont souvent acceptés avec un contrôle minimal, tandis que les assertions surprenantes qui vont à l'encontre de la sagesse conventionnelle sont accueillies avec un scepticisme prononcé, même si les arguments s'avèrent finalement corrects lors d'une lecture ligne par ligne.

4. Nous laissons ici la notion de mathématiques dans une certaine imprécision. On peut adopter une approche prescriptive, par exemple en utilisant la définition de Davis et Hersh [12] qui définissent les mathématiques comme "l'étude des objets mentaux dotés de propriétés reproductibles". Ou bien on peut adopter une approche descriptive, selon laquelle les mathématiques sont l'activité que
les mathématiciens pratiquent concrètement. Notre propos penche plutôt pour cette dernière approche.

4.2. Le test de l'odeur.

Jusqu'à présent, cet état de fait a été raisonnablement satisfaisant ; les mathématiciens humains qui suivent de bonnes heuristiques et intuitions ont tendance à produire des preuves convaincantes et largement correctes, la plupart des erreurs étant réparables, tandis que les mathématiciens qui manquent d'une telle intuition ont tendance à produire des preuves contenant suffisamment de problèmes superficiels pour que l'on puisse légitimement se méfier du contenu avant même de l'avoir vérifié attentivement. De manière informelle, les arguments mathématiques générés par l'homme ont tendance à être accompagnés [5] d'une "odeur" que les mathématiciens expérimentés utilisent (peut-être subconsciemment) pour obtenir leurs premières impressions sur le caractère convaincant de l'argument, bien avant d'avoir pu vérifier les étapes individuelles de cet argument. Par exemple, l'article de blog "Ten signs a claimed mathematical breakthrough is wrong" d'Aaronson [15] énumère quelques exemples courants d'arguments qui dégagent une telle "mauvaise odeur", détectable bien avant d'avoir localisé un défaut logique spécifique dans l'argument proposé. Et toutes les erreurs ne sont pas également désastreuses ; certaines erreurs peuvent même avoir une certaine valeur bénéfique, par exemple en révélant une approche prometteuse avant de pouvoir la valider pleinement [16].

Une composante d'une "bonne odeur", comme l'a noté [6] Thurston [18], est le sentiment qu'un argument apporte de la compréhension ou de la perspicacité ; qu'il ne se contente pas de montrer qu'un certain ensemble d'hypothèses entraîne logiquement une conclusion donnée, mais qu'il fournit un récit causal expliquant comment cette implication était possible, et quelles parties de l'argument effectuaient le "travail lourd", quelles parties étaient nouvelles ou surprenantes par rapport à la littérature antérieure, et lesquelles étaient des considérations techniques de routine. Ces interprétations et impressions du texte mathématique ne sont généralement pas capturées dans les cadres officiels des mathématiques rigoureuses, comme la logique du premier ordre ou la théorie des ensembles ; mais elles sont essentielles pour permettre aux mathématiciens lisant l'argument d'en tirer des leçons plus larges sur la manière dont on pourrait s'attendre à ce que les arguments se généralisent à d'autres contextes, ou interagissent avec d'autres méthodes de la littérature. Ces structures narratives aident également à renforcer la confiance dans la robustesse d'un argument ; une erreur de signe isolée dans un calcul pourrait invalider un long argument mathématique, mais si la preuve avait une stratégie claire sur la manière dont les difficultés clés de l'argument étaient systématiquement isolées et traitées, en suivant des analogies avec des arguments réussis antérieurs dans la littérature, il devient alors plus probable que les erreurs locales dans l'argument puissent être réparées tout en restant fidèles à l'esprit de la preuve originale.

5. Nous utilisons cette métaphore sensorielle par analogie avec le concept d'"odeur de code" en génie logiciel.

6. Voir aussi l'article [17] du deuxième auteur, qui soutient que les "bonnes" mathématiques, quelle que soit leur définition initiale, tendent souvent en pratique à s'intégrer dans des récits mathématiques plus larges, tels que la dichotomie entre structure et aléatoire, ou la capacité de l'algèbre à explorer des questions sur la géométrie (ou vice versa).

4.3. La formalisation à la rescousse ?

Plusieurs développements pourraient forcer la communauté mathématique à réévaluer ce standard de preuve semi-formel. L'un d'eux est de nature technique : à mesure que les mathématiques ont mûri et sont devenues plus sophistiquées (et de plus en plus assistées par ordinateur), les arguments sont devenus plus longs et plus complexes, les articles de pointe dans certains domaines dépassant couramment la centaine de pages, rendant la vérification ligne par ligne par les relecteurs humains de plus en plus ardue. En pratique, cela a signifié qu'un tel contrôle minutieux n'est pas toujours effectué, sauf pour les résultats les plus médiatiques et importants, conduisant à une confiance croissante (excessive) dans le sens de "l'odeur" susmentionné pour évaluer la crédibilité des arguments mathématiques.

Il semble possible que ces problèmes puissent être résolus (ou du moins atténués) par des moyens techniques, et en particulier par le déploiement plus généralisé d'assistants de preuve formels (comme `lean` ou `rocq`) qui peuvent vérifier automatiquement la validité d'un argument mathématique s'il est écrit dans un certain langage informatique précis [19]. Une telle formalisation reste trop fastidieuse à l'heure actuelle pour être déployée systématiquement (la conversion d'une preuve traditionnelle, rédigée de manière informelle, dans un tel langage formel prend généralement de cinq à dix fois plus de temps que la rédaction de cette preuve elle-même), mais des efforts significatifs sont en cours pour rendre le processus plus rapide et plus convivial, par exemple en intégrant des outils d'ia pour réaliser une "auto-formalisation" partielle (ou éventuellement complète) [20].

4.4. Limites de la vérification formelle.

Mais même si ces problèmes techniques sont résolus, et que les preuves mathématiques sont systématiquement accompagnées d'une vérification formelle d'exactitude, plusieurs nouveaux problèmes surgissent, en particulier dans un avenir proche où des arguments de plus en plus sophistiqués pourraient être partiellement ou entièrement générés par des outils d'ia. Premièrement, la vérification formelle ne certifie pas qu'un argument formalisé établit un énoncé mathématique formel, mais n'exclut pas les erreurs de traduction entre l'énoncé formel et l'énoncé original visé. Par exemple, le dernier théorème de Fermat affirme que pour tout nombre naturel n supérieur à 2, il n'existe pas de solutions en nombres naturels a, b, c à l'équation $a^n + b^n = c^n$; mais implicitement dans cette description informelle, il y a la convention que les nombres naturels commencent à 1 plutôt qu'à 0. Une ia chargée de résoudre ce problème pourrait supposer à tort que a, b, c sont autorisés à être nuls, et sur cette base produire une preuve (formellement certifiée) que le dernier théorème de Fermat est faux ! Ainsi, bien que la formalisation puisse en principe réduire considérablement le besoin d'examen humain attentif des textes mathématiques informels, elle n'élimine pas entièrement ce besoin [7].

7. Il est même théoriquement possible que les mathématiques elles-mêmes puissent être "piratées" en manipulant subtilement la formalisation des définitions clés dans les bibliothèques mathématiques formelles standard telles que mathlib ; voir [21]. Paradoxalement, la recherche mathématique de plus en plus collaborative, sociale et à grande échelle, bien qu'étant généralement un développement très positif, peut également accroître les vulnérabilités potentielles à de telles attaques qui n'étaient pas une préoccupation majeure aux époques précédentes où les mathématiques étaient principalement pratiquées par
de petits groupes d'individus.

Deuxièmement, même dans le cadre purement abstrait des mathématiques avancées, seule une partie d'un argument donné peut être formulée dans le type de logique déductive qui se prête à la formalisation. Bien que la preuve déductive reste le noyau crucial de la plupart des travaux mathématiques, il existe une pénombre de raisonnement heuristique, empirique ou métamathématique autour de ce noyau, qui fournit des informations précieuses sur les raisons pour lesquelles l'argument fonctionne, s'il s'étend à d'autres contextes, quelle est la motivation pour poursuivre ces questions, et comment on pourrait reconstruire l'argument à partir de principes plus fondamentaux. Les preuves rédigées par des humains, de par leur nature, ont tendance à fournir cette pénombre organiquement dans le processus d'écriture (en particulier si les auteurs sont habiles à l'exposition) ; mais une ia qui a été entraînée spécifiquement sur le critère de la correction formelle, au détriment de toute autre considération, pourrait produire des preuves "inodores" qui ressemblent superficiellement à une preuve humaine bien rédigée, et pourraient même réussir les tests de vérification formelle, mais qui restent étrangement insatisfaisantes - remplissant à la lettre l'objectif explicite d'établir une affirmation mathématique donnée, tout en offrant beaucoup moins de compréhension que prévu sur le champ mathématique plus large dont cette affirmation fait partie. Dans un monde où tous les médias générés sont polis par l'ia pour atteindre un éclat élevé, y compris les preuves mathématiques avec une belle composition et des explications claires produites par GPT, quelque chose est-il perdu en abandonnant le monde plus sale et plus désordonné du texte manuscrit (ou du moins tapé à la main) ?

4.5. Adaptation aux défis antérieurs.

La communauté mathématique s'est adaptée aux défis technologiques antérieurs posés à son standard de preuve. Les grandes preuves assistées par ordinateur, comme la preuve du théorème des quatre couleurs [22] ou la conjecture de Kepler [23], ont d'abord suscité des controverses, étant difficiles à vérifier entièrement à la main ; mais avec le temps, de nouveaux standards pour établir la confiance dans ces types d'arguments ont été établis, comme la fourniture de codes reproductibles, l'isolation des composantes computationnelles d'un argument dans des lemmes spécifiques et clairement énoncés, séparés des aspects plus conceptuels d'un article, et la fourniture de données supplémentaires et de "sommes de contrôle" pour vérifier que les calculs générés par ordinateur concordent avec divers "tests de cohérence". En effet, ces développements ont déplacé les standards de preuve en mathématiques vers ceux des sciences naturelles, où les arguments théoriques et les expériences empiriques, lorsqu'ils sont correctement conçus, exécutés et rapportés, sont des sources acceptables de vérité scientifique.

4.6. L'évolution des mathématiques assistées par l'ia.

Des évolutions similaires auront lieu [8] avec l'avènement des mathématiques significativement assistées ou générées par l'ia. La charge de produire des preuves déductives vérifiées pourrait de plus en plus incomber aux ordinateurs plutôt qu'aux humains, les preuves étant de plus en plus restruc-

8. Notre réflexion ici a été influencée par les points de vue d'autres mathématiciens sur ce sujet, notamment [24], [25], [16], [26], ainsi que par la discussion plus large dans [27], [28].

turées[9] pour que les calculs fastidieux qui seraient auparavant soigneusement agencés pour être vérifiables par l'homme soient de plus en plus externalisés vers des outils logiciels. Par exemple, les phrases infâmes en mathématiques telles que "la preuve est laissée au lecteur" ou "Par des arguments standards, nous avons" pourraient être remplacées par un appel à un llm qui produit à la fois des justifications lisibles par l'homme et vérifiables par ordinateur pour de telles affirmations. Avec les progrès de l'auto-formalisation, il deviendra également beaucoup plus facile d'explorer comment un argument donné change par rapport à des choix spécifiques de fondements des mathématiques, permettant à la métamathématique[10] d'un résultat d'être rigoureusement discutée et explorée simultanément avec le résultat mathématique lui-même.

En même temps, une attention et un intérêt accrus pourraient être accordés à l'avenir par les mathématiciens humains aux aspects "plus doux" du raisonnement mathématique, tels que les heuristiques et la motivation pour poursuivre un résultat ou sélectionner une stratégie de preuve pour ce résultat, les preuves expérimentales[11] en faveur (ou contre) le résultat, ou le processus d'essais et d'erreurs menant à la découverte d'un argument fonctionnel. Ces aspects ne sont pas aussi faciles à vérifier et mesurer automatiquement que la preuve déductive, et donc moins adaptés[12] aux stratégies d'apprentissage automatique comme l'apprentissage par renforcement. Il est concevable que les mathématiciens professionnels adoptent de plus en plus[13] des modes d'argumentation provenant d'autres disciplines, comme les sciences théoriques et expérimentales ou même les sciences humaines, pour étayer leurs arguments déductifs centraux par des types de raisonnement supplémentaires, tels que l'analyse statistique de données expérimentales, ou la théorisation spéculative guidée à la fois par des résultats mathématiques confirmés et des principes philosophiques non rigoureux. Historiquement[14], les mathématiciens ont été réticents à s'éloigner trop de leur "étalon-or" de la preuve déductive rigoureuse, en raison en partie des nombreux exemples visibles

9. Pour plusieurs exemples concrets de telles restructurations et une exploration plus approfondie de ces développements, voir [29].

10. Un exemple de métamathématiques est la démonstration inverse (voir, par exemple, [30]) d'un théorème, qui vise à déterminer quels axiomes mathématiques (par exemple, l'axiome du choix ou le principe du tiers exclu) sont réellement nécessaires pour établir un résultat donné. Traditionnellement, la démonstration inverse d'un résultat n'est explorée que de nombreuses années après la démonstration originale de ce résultat, et requiert une formation spécialisée en logique ainsi qu'une expertise du sous-domaine mathématique auquel appartient le théorème.

11. En particulier, étant donné la capacité croissante de l'ia à "deviner" la réponse à des questions mathématiques même extrêmement complexes sans disposer de la moindre preuve formelle, il deviendra de plus en plus nécessaire de développer des procédures standard pour citer et intégrer de manière responsable ces conjectures non vérifiées dans la littérature mathématique.

12. Un autre obstacle à l'automatisation de ces aspects du processus de recherche mathématique est le manque relatif de données ; la littérature publiée a tendance à se concentrer sur les démonstrations réussies des résultats, au détriment de la description détaillée des processus (souvent très riches et nuancés), formels et informels, qui ont conduit à ces démonstrations.

13. On peut notamment envisager une division croissante du travail dans l'avenir de la recherche mathématique : si tous les mathématiciens doivent posséder une connaissance générale des différentes étapes de la proposition, de l'établissement et de l'interprétation des résultats mathématiques, chaque mathématicien pourra se spécialiser de plus en plus dans quelques aspects de ce processus, par exemple en se concentrant sur l'utilisation d'assistants d'ia pour démontrer des résultats sous la direction d'un membre plus expérimenté du groupe de recherche, ou sur l'utilisation des publications les plus récentes, produites conjointement par des mathématiciens et des assistants d'ia, pour proposer de nouvelles pistes de recherche.

14. Par exemple, une proposition antérieure de Jaffe et Quinn [31] visant à développer systématiquement un domaine de "mathématiques théoriques" a reçu un accueil largement négatif de la part des mathématiciens professionnels, ce qui a conduit à de multiples répliques, dont l'article susmentionné de Thurston [18].

de mathématiques de faible qualité qui peuvent être produites lorsque l'on ne respecte plus de tels standards [15] ; mais dans une ère future où les preuves peuvent être automatiquement générées et vérifiées de manière hautement fiable, il pourrait y avoir plus d'opportunités d'explorer en toute sécurité de tels modes plus larges de raisonnement et de discussion mathématiques.

Ces nouvelles technologies pourraient également avoir un impact négatif significatif sur les objectifs à long terme des mathématiques. Au niveau éducatif, nous voyons déjà de nombreux étudiants qui recourent presque immédiatement aux outils d'ia modernes pour effectuer leurs travaux scolaires assignés, atteignant l'objectif immédiat de produire des réponses vérifiables à un problème donné au détriment du développement de compétences mathématiques plus durables et de l'intuition ; de même au niveau de la recherche, le "quatrième paradigme" des mathématiques pilotées par les données [33] pourrait concevablement être si réussi qu'il évince les paradigmes plus traditionnels des preuves empiriques, du raisonnement théorique et de la numérisation computationnelle (le second étant actuellement le paradigme dominant pour les mathématiques pures), ainsi que la grande valeur que les mathématiciens humains [16] tirent de l'intuition visuelle, kinesthésique et sensorielle, ou de l'intuition ancrée dans notre familiarité avec les lois de la physique, de l'économie, de la biologie, etc. Même en supposant une mise en œuvre totalement fiable des méthodes formelles, une adoption non critique de l'assistance de l'ia dans le domaine de la recherche mathématique pourrait conduire à la conséquence indésirable d'une inondation [17] d'articles en grande partie générés par l'ia contenant des résultats techniquement corrects et nouveaux, mais qui ne contribuent pas aux récits mathématiques plus larges et ne construisent pas d'intuition ni pour les auteurs ni pour les lecteurs. Les impressions négatives produites par un tel travail de faible qualité pourraient conduire à une stigmatisation contre l'utilisation de l'assistance de l'ia, même la plus prudente et responsable, dans la recherche mathématique.

15. Kim [32] invoque une métaphore monétaire pour décrire la dynamique sociale : les mathématiciens professionnels ont besoin d'accumuler une certaine "monnaie" de crédibilité, en prouvant de nouveaux résultats mathématiques difficiles, avant de pouvoir "se permettre" de "dépenser" cette monnaie dans des activités spéculatives, telles que la formulation de conjectures ou la philosophie sur les conséquences plus larges d'un résultat.

16. Dans un registre similaire, des notions esthétiques telles que la "beauté" ou l'"élégance" d'un raisonnement mathématique pourraient se dissocier encore davantage qu'elles ne le sont actuellement de la validité formelle de ces raisonnements. Prenons par exemple les démonstrations générées par `alphaproof` [34] pour des problèmes de l'Olympiade mathématique internationale de 2024, qui contenaient de nombreuses étapes redondantes ou inexplicables, mais qui ont néanmoins été formellement vérifiées comme étant des solutions correctes. Voir également la discussion dans [25].

17. On pourrait y voir une illustration de la loi des conséquences imprévues. Aux époques mathématiques passées, où l'obtention d'une démonstration mathématique rigoureuse exigeait un effort humain considérable, l'activité mathématique se concentrait naturellement sur les problèmes jugés d'intérêt par la communauté mathématique, même si la question philosophique de la véritable signification d'un résultat donné était souvent négligée par la plupart de ses membres. L'évolution de la littérature était suffisamment lente pour que ce mécanisme largement social de détermination de la signification mathématique puisse se corriger de lui-même au fil du temps. Dans une ère future où les résultats mathématiques pourront être produits en masse à des vitesses considérablement plus rapides grâce à l'automatisation, de telles questions philosophiques pourraient nécessiter une attention beaucoup plus soutenue. Voir également [35] et [36] sur la nécessité de porter des jugements de valeur, notamment sur la fiabilité de l'auteur, lorsqu'il s'agit de décider d'accorder ou non de l'attention à un résultat mathématique revendiqué.

4.7. Application des questions philosophiques à l'utilisation réelle de l'ia.

Tout contenu qui est une référence fondamentale pour d'autres recherches porte des responsabilités supplémentaires, et les mathématiques ne font pas exception. Nous pouvons certifier formellement la validité de tout argument mathématique généré par l'ia ; mais la validité n'est qu'une composante de la valeur, et il existe des jugements de valeur nuancés qui sont nécessaires pour présenter la recherche pilotée par l'ia dans des situations réelles. Quels éléments de l'ensemble potentiellement vaste de résultats triviaux et non triviaux le chercheur trouve-t-il particulièrement intéressants et dignes d'être partagés au sein et au-delà du domaine de recherche, et comment ce matériel est présenté à un public plus large, n'ont pas été standardisés parmi les chercheurs humains. Il existe également des incertitudes sur la manière dont la priorité et le crédit sont attribués. La recherche assistée par l'ia présente également de nouvelles ramifications éthiques et juridiques et des questions encore sans réponse sur les droits de propriété intellectuelle du contenu généré par l'ia (y compris les preuves).

Quels principes devraient guider les chercheurs dans la décision de la pertinence et de la meilleure application d'un modèle d'ia ou d'un autre, ou si l'ia est un bon choix en général ? Dans les domaines académiques, il n'est pas déraisonnable de supposer que la plupart de ceux qui poursuivent la voie de la recherche académique le font par désir de rendre le monde meilleur et d'y apporter des contributions significatives. Les mathématiciens voudront prioriser les cas d'usage les plus bénéfiques pour les mathématiques. Les chercheurs de tous les domaines voudront fréquemment prioriser non seulement les utilisations qui bénéficient à leur propre domaine mais qui ont également des avantages interdisciplinaires. Et on peut considérer que la plupart de ceux qui utilisent l'ia à des fins de recherche voudront prioriser les utilisations qui bénéficient à l'humanité plutôt que celles qui lui nuisent. Il est par conséquent important que, dans le domaine du développement de l'ia, il soit mis en évidence qui bénéficie de ces outils et quels sont les avantages qui se produisent, pour aider les gens à identifier comment optimiser de manière responsable les résultats autant que possible.

4.8. Propriété intellectuelle et responsabilité.

La question de la propriété intellectuelle et de la responsabilité (ou peut-être, de la redevabilité) est à elle seule un champ de mines et nécessite une discussion attentive. Lorsque l'ia est appliquée à un problème, qui est responsable des erreurs ? Qui obtient le crédit pour les idées ? Il peut ne pas s'agir de la même partie, et il peut s'agir de parties qui ne sont pas clairement définies. Jusqu'à présent, une grande partie de l'accumulation de données d'entraînement pour les grands modèles de langage (llm) a été argumentée (par leurs développeurs) comme relevant de la doctrine de l'"utilisation équitable" (fair use). Aux États-Unis, l'application de l'"utilisation équitable" a une certaine flexibilité en fonction (entre autres) de l'objectif d'utilisation de l'intelligence personnelle [37]. Comme expérience de pensée, nous pouvons considérer si un plus grand bénéfice justifie une plus grande utilisation [38].

Serait-il raisonnable de prétendre qu'il est d'utilisation équitable de puiser dans toutes les connaissances enregistrées dans une situation où il est prévu de sauver le monde d'une calamité imminente ? Une application aussi large s'appliquerait-elle encore s'il s'agissait de sauver le monde d'une menace existentielle plus lointaine (par exemple, le changement climatique) ? Qu'en est-il s'il s'agissait

“seulement” de mettre fin à toutes les maladies ? Ou simplement d’éradiquer le cancer ? Comme toutes ces applications sont considérées comme des applications bénéfiques de l’ia, est-il alors raisonnable d’accorder à l’ia l’utilisation de toutes les informations enregistrées pour rendre de telles merveilles possibles ?

Au-delà de l’argument problématique pour une interprétation extrêmement large de l’“utilisation équitable”, des normes et protocoles clairs pour l’attribution du crédit et la citation sont désespérément nécessaires. Les cas d’utilisation de l’ia puiseront non seulement dans les données du chercheur, mais aussi dans les informations sur lesquelles l’ia a été précédemment entraînée, les choix des informations sur lesquelles l’ia a été entraînée (effectués par des ingénieurs logiciels et des concepteurs qui peuvent n’avoir aucune interaction avec le chercheur principal) et, bien sûr, les contributions de l’ia elles-mêmes. Le système de citation académique traditionnel est-il adéquat pour attribuer un crédit approprié dans une situation impliquant potentiellement des centaines ou des milliers de contributeurs “cachés”, ou est-il adéquat de simplement citer le modèle d’ia lui-même ? L’utilisation non divulguée de l’ia pour effectuer une partie significative de la rédaction d’articles de recherche a provoqué des réactions particulièrement vives, de nombreux universitaires considérant une telle pratique comme comparable au plagiat ; ironiquement, cela a conduit certains chercheurs qui tirent profit de leurs outils à dissimuler encore plus leur utilisation. Il est clair que de nouvelles normes professionnelles et pratiques concernant la divulgation et l’utilisation de l’ia devront être développées [18].

L’ia est également sur le point de créer des boucles de citation circulaires potentiellement généralisées, un processus humoristiquement baptisé “citogenèse” par Randall Munroe [19] en 2001. Par exemple, suite au succès récent des outils de “recherche approfondie” (deep research) par l’ia [40] pour découvrir des solutions à des problèmes ouverts enfouis dans la littérature obscure, le deuxième auteur a contribué à lancer un effort sur un site de problèmes mathématiques ouverts [41] pour utiliser systématiquement ces outils afin de rapporter la littérature connue sur ces problèmes, ou l’absence d’une telle littérature. Bien que cela ait ajouté une réelle valeur au site, nous avons également constaté que les outils de recherche approfondie utilisaient ces rapports comme une source faisant autorité pour leurs recherches, avec la conséquence involontaire que la synthèse de ces recherches sur le site interférait avec toute utilisation ultérieure de ces outils pour découvrir une littérature véritablement nouvelle sur de tels problèmes ! Ainsi, même en l’absence d’intention malveillante, la puissance croissante de ces outils nécessite une vérification plus approfondie de la provenance des informations citées.

5. Coûts et avantages de l’ia

5.1. Impacts économiques et sociétaux : qui en bénéficie ?

Compte tenu de l’impact déjà significatif de l’ia sur les individus, ainsi que de son rythme de développement rapide, il est trop facile d’entrevoir une voie dans laquelle l’ia monte en puissance pour représenter une menace existentielle pour l’espèce. À chaque pas en avant, les développeurs

18. Pour une première discussion de ce sujet par l’un des auteurs, voir [39].
19. https://xkcd.com/978/.

et autres personnes influentes doivent examiner attentivement qui bénéficie de ces avancées et qui en est lésé. Nous proposons que tout développement ultérieur devrait prioriser le bénéfice pour l'humanité dans son ensemble et que les applications de l'ia devraient rester directement utiles aux humains (individuellement ou collectivement).

Pour chaque cas d'usage individuel, une évaluation devrait être faite pour articuler qui sont les bénéficiaires visés. Ce modèle d'ia spécifique ou cette mise en œuvre d'un modèle bénéficiera-t-il à la société dans son ensemble ou ne fournira-t-il des avantages tangibles (tels que des économies de coûts) qu'à un petit groupe d'individus ? Les outils d'ia sont d'une puissance et d'une complexité telles qu'un gain économique extrême pour un petit nombre d'individus qui lèserait des millions de personnes représente un coût moral intolérable et inacceptable. Nous devons faciliter des mises en œuvre de l'ia qui préservent et qui valorisent l'humanité des humains plutôt que de les considérer comme des marchandises.

Nous n'avons pas besoin de chercher loin pour trouver des résultats désastreux de la priorisation du capital sur le bien-être humain. Souvent caractérisés comme arbitrairement anti-technologie et anti-progrès, les ouvriers du textile de Nottingham du début du XIXe siècle qui se faisaient appeler "Luddites" se sont violemment opposés à l'automatisation qui les dépossédait de leurs emplois et les remplaçait par des travailleurs moins qualifiés et moins bien payés. La menace immédiate pour leurs emplois représentait une menace existentielle pour leurs moyens de subsistance dans un climat économique difficile marqué par un chômage élevé et une inflation galopante. Bien que nous considérions a posteriori l'automatisation de la révolution industrielle comme généralement bénéfique pour la société, ces avantages sont venus avec des coûts humains réels et mesurables.

Aujourd'hui, contrairement à l'époque des Luddites, nous voyons déjà des travailleurs qualifiés remplacés non par une main-d'œuvre humaine moins chère, mais par l'ia. Les emplois de niveau débutant ont historiquement été la voie vers la prospérité financière et sociale pour une génération naissante de travailleurs. Lorsqu'ils disparaissent simplement, l'opportunité disparaît avec eux. Le désespoir et le ressentiment s'accumulent jusqu'à la colère et l'indignation, tandis que les humains se placent en opposition directe avec les outils qui promettaient d'améliorer leur qualité de vie.

Même si toutes les technologies émergentes ont des avantages pour l'humanité dans son ensemble, elles s'accompagnent également d'un coût humain réel. Pour une technologie radicalement disruptive comme l'ia, les coûts humains doivent être quantifiés au niveau local et mondial et soigneusement mis en balance avec les avantages. Les mesures que nous utilisons pour cette évaluation sont encore floues et mal définies. Devons-nous continuer comme avant, à considérer seulement les gains et pertes monétaires ? Devrions-nous considérer l'accès accru aux ressources mis en balance avec les ressources perdues ? Prenons-nous suffisamment en compte les avantages plus intangibles de la qualité de vie et du bonheur, et si oui, comment comparer ces intangibles avec des gains plus quantitatifs ?

Le climat économique actuel cherche malheureusement une "arme miracle" (*Wunderwaffe*) qui est optimisée pour la puissance et les impacts les plus larges possibles, dans l'espoir qu'elle pourra distancer tout problème potentiel. Mais une incapacité à quantifier le coût humain de nos technologies émergentes fait un grand tort à l'humanité dans son ensemble pour le bénéfice d'une minorité

sélectionnée. De plus, le climat actuel où l'ia est mise en œuvre simultanément dans pratiquement toutes les sphères de la société, sans considération de savoir si elle apporte un avantage significatif aux utilisateurs finaux, ne fait qu'aliéner et frustrer les gens de tous les horizons. Nous voyons déjà la réaction naturelle à une technologie imposée aux individus sans consentement - ressentant une perte de contrôle, leur premier instinct est de rejeter toutes les technologies de l'ia, même au risque de jeter le "bébé" (les usages de l'ia qui offrent des avantages quantifiables dans leur vie) avec l'eau du bain. Si nous pouvons au contraire garder notre technologie concentrée en premier lieu sur l'amélioration quantifiable de la plupart ou de toutes les vies humaines, nous serons beaucoup moins susceptibles de nous détruire que si l'objectif unique de ces technologies est la marchandisation du travail mécanique, du travail numérique et du travail humain.

5.2. Évaluation des coûts de l'ia

Parallèlement aux coûts humains directs, aucune mise en œuvre éthique de l'ia ne peut avoir lieu sans examiner d'autres coûts plus opaques et cachés. Le coût le plus substantiel et le plus immédiatement apparent du développement et de la construction d'une infrastructure d'ia efficace provient de la réalité que ces technologies, contrairement à celles de la révolution informatique des années 1970, ne peuvent pas être développées comme un passe-temps ou une industrie artisanale - il n'y a pas de garage de pièces informatiques qu'un penseur innovant comme Steve Jobs pourrait utiliser pour construire un empire. Les modèles d'ia qui ont été construits nécessitent un investissement massif en matériel, serveurs, talents et pré-entraînement bien avant d'arriver à une ia fonctionnelle, et encore moins rentable.

Une meilleure comparaison pour l'échelle requise par l'ia est le réseau ferroviaire transcontinental construit aux États-Unis dans la seconde moitié du XIXe siècle. Les entreprises qui ont construit les chemins de fer ont dû développer et construire une flotte de locomotives massives, et planifier et poser des milliers de kilomètres de voies avant que le premier train puisse transporter rapidement et fiablement des marchandises de l'Iowa à San Francisco, débloquant ainsi les rendements économiques sur lesquels ces entreprises avaient parié.

La dépense initiale énorme pour les technologies basées sur l'ia a conduit les développeurs à poursuivre un modèle capitaliste axé sur le profit, créant une nouvelle classe d'élites technologiques qui manient d'énormes sommes de capital investi et de dette gérée tout en manœuvrant stratégiquement pour capturer et détenir des ressources finies (en terres, énergie, eau, main-d'œuvre qualifiée, etc.), tout comme les barons voleurs de l'âge d'or du XIXe siècle. Comme à cette époque, l'échelle de ces investissements a entraîné des inégalités massives en matière de stabilité économique, d'accès à ces technologies et de qualité de vie générale dans le monde développé.

Notre société a déjà commencé à reconnaître les coûts environnementaux significatifs qu'exige l'ia à grande échelle. Une consommation lourde d'énergie et d'eau crée des défis quotidiens importants pour ceux qui vivent dans l'ombre des installations étendues que ces modèles d'ia exigent. Il a été sérieusement suggéré (voir, par exemple, [42]) que des solutions générées par l'ia pourraient être appliquées pour atténuer ou éliminer les lourds coûts climatiques de deux siècles d'utilisation de la technologie humaine. Et peut-être que les coûts marginaux d'exploitation de ces outils diminueront

avec le temps à mesure que l'infrastructure sera construite et que des utilisations plus efficaces du calcul seront développées. Cependant, à ce jour, aucun des grands modèles d'ia en opération n'a fourni une solution pour compenser ne serait-ce que leur propre consommation de ressources et leurs émissions de déchets.

De plus, il est à noter que les outils d'ia modernes ne recherchent ni n'intuitent la "vérité" par une manifestation dans le monde physique, ou une compréhension de la nature immuable des lois physiques de notre réalité ; au lieu de cela, ces modèles s'appuient fortement sur des données générées par l'homme, souvent sans attribution, ainsi que sur des quantités importantes de retours humains pour s'améliorer itérativement. Les modèles ne peuvent pas être construits pour être moins dépendants du travail intellectuel humain sans un risque sérieux de contaminer notre corpus collectif d'informations avec des informations générées par l'ia. Il y a une limite claire à la quantité d'ia qui peut être utilisée pour générer de "nouvelles informations" dans un domaine avant que l'effondrement de l'ia [43] ne devienne un problème sérieux. Sans une quantité suffisante de contenu authentique, l'ia devient déconnectée de la réalité, prise dans un mode de pensée qui n'a aucun lien avec le monde réel et qui entrave considérablement les interactions significatives à l'interface humain/ia. Les mathématiques, avec leur processus de vérification formelle, peuvent avoir une tolérance plus élevée à la contamination par l'ia que d'autres domaines ; mais comme nous l'avons vu, elles ne sont pas complètement immunisées contre ce danger.

5.3. La fracture numérique.

Un coût social supplémentaire significatif à considérer est le potentiel des technologies de l'ia à exacerber les inégalités existantes ou à en créer de nouvelles. En principe, tous les humains ont la capacité d'utiliser leurs talents intellectuels naturels (en supposant une éducation adéquate et un environnement favorable, bien sûr) ; mais les tendances dans l'application des modèles d'ia de pointe montrent déjà que les outils d'ia à grande échelle peuvent n'être disponibles que pour des groupes de recherche bien financés ou bien connectés, ou pour des individus les plus disposés à céder leurs données personnelles et à passer outre les préoccupations éthiques concernant l'utilisation de tels modèles. Cela crée une "fracture numérique" fondamentale entre les "ayant" et les "non-ayant" de l'ia.

Il est primordial de prioriser un accès équitable lorsque l'ia a la capacité d'améliorer radicalement les performances de recherche. Cependant, dans le paysage actuel de l'ia, une deuxième fracture numérique, plus nuancée, apparaît. Lorsque les modèles d'ia dominants sont capitalisés, privatisés et en concurrence pour des ressources finies (investissement et base d'utilisateurs dépendante), ils sont (peut-être involontairement) incités à développer des capacités "pointues" pour conserver un avantage concurrentiel les uns sur les autres, plutôt qu'à fournir des performances cohérentes et uniformes dans différents domaines. Comme les individus sont enfermés dans un modèle plutôt qu'un autre en raison de négociations institutionnelles et de contraintes de marché, nous devons considérer le risque qu'un modèle donne un avantage significatif sur un autre dans une sphère de recherche particulière, créant des divisions même au sein du sous-groupe qui a un accès fiable et facile aux ressources d'ia.

D'un autre côté, bon nombre des avantages des modèles d'ia dans la recherche scientifique et en sciences humaines ne nécessitent pas nécessairement les modèles les plus avancés. Des "modèles locaux" plus petits, ainsi que des technologies non-llm comme les assistants de preuve, démontrent la capacité à retourner des résultats significatifs plus rapidement et plus efficacement que les modèles qui nécessitent des centres de données massifs traitant la somme de toutes les connaissances humaines. Il y a un potentiel significatif pour pouvoir distiller des modèles plus petits à partir des plus grands existants afin de tirer parti des capacités d'ia les plus avancées avec de petites bibliothèques d'entraînement définies par l'utilisateur et ciblées soigneusement sur le domaine de recherche spécifique. Peut-être qu'un ensemble diversifié de modèles plus petits et plus ciblés, maintenus par une communauté d'utilisateurs, pourrait émerger comme une alternative viable aux modèles extrêmement grands et coûteux disponibles aujourd'hui. Un soutien accru pour de tels projets communautaires pourrait aider à atténuer le problème d'accès inéquitable.

Bien que bon nombre de ces projets plus petits puissent être développés et exécutés par des institutions publiques et privées à plus petite échelle, les praticiens de l'industrie et les décideurs politiques ont appelé à des actions réglementaires pour créer et préserver un accès équitable aux technologies de l'ia [44]. Dans le cadre de cet effort, il y aurait des avantages significatifs à investir dans le développement d'une coalition nationale ou multinationale de recherche avancée en ia et dans le développement d'une ressource (ou de modèles) d'ia large, financée publiquement et accessible au public [45], pour apporter un accès à l'ia aux individus et aux groupes qui seraient autrement laissés pour compte par les modèles privés et corporatisés qui dominent actuellement le domaine.

5.4. Réduction des dommages.

Aux débuts de l'aviation, le transport aérien était une technologie incroyablement dangereuse, avec d'innombrables accidents horribles. Aujourd'hui, c'est le mode de transport le plus sûr et le plus fiable sur de longues distances. Tout comme l'ia a le potentiel de conduire à des résultats catastrophiques à court terme, pour qu'elle suive une trajectoire similaire (avec, espérons-le, moins d'incidents mortels), des actions décisives seront nécessaires pour réduire les dommages. Des meilleures pratiques doivent être définies [46] et une formation et une réglementation conçues pour renforcer les utilisations les plus responsables de la technologie, tout en décourageant ou en interdisant les utilisations dissimulées ou nuisibles.

C'est une tâche qui nécessite une grande concentration (*a tough needle to thread* : enfiler du fil dans une aiguille avec un châs très étroit). D'une part, un individu qui utilise l'assistance de l'ia avec prudence et responsabilité pourrait être dépassé à court terme par des rivaux moins scrupuleux qui utilisent des pratiques d'ia plus rapides, mais moins fiables, pour accélérer leur travail. En même temps, ces mêmes individus pourraient être moqués, condamnés et exclus par des cohortes de pairs méfiants envers l'ia, simplement pour avoir osé envisager la possibilité d'incorporer la technologie dans les flux de travail de leur profession. L'approche actuelle découlant d'un large laisser-faire, qui permet à la technologie de l'ia de se développer à un rythme non contrôlé, ne semble pas prometteuse pour qu'une approche aussi nuancée et responsable de l'adoption prévale.

Il existe quelques précédents sur lesquels s'appuyer pour obtenir des conseils. Le développement

rapide de wikipédia au début du XXIe siècle a d'abord causé des perturbations dans les systèmes éducatifs, car de nombreux étudiants ont commencé à incorporer aveuglément du texte de cette ressource en ligne mot pour mot dans leurs devoirs, et de nombreux enseignants ont réagi en interdisant l'utilisation de cette ressource encyclopédique. Les critiques sur le manque de fiabilité et les biais potentiels de wikipédia étaient courantes. Cependant, à mesure que le site a mûri et que le monde universitaire s'est familiarisé avec ses forces et ses faiblesses, un consensus approximatif a émergé sur la manière d'incorporer cette ressource dans l'éducation et la recherche. Il est devenu encouragé, ou du moins toléré, pour les étudiants et les chercheurs d'utiliser wikipédia comme point de départ pour une enquête sur un sujet donné ; et, au lieu d'utiliser directement son texte, les étudiants sont invités à suivre les sources secondaires fournies par le site, ou à les vérifier par rapport à des sources d'information indépendantes. Aujourd'hui, wikipédia est largement accepté comme une ressource utile dans le monde universitaire.

Pourrions-nous atteindre un niveau similaire d'acceptation responsable avec l'ia ? Nous sommes prudemment optimistes quant au fait que cela soit possible ; mais cela nécessitera un effort soutenu et des orientations philosophiques claires. Par exemple, nous croyons qu'il y a un impératif moral et éthique à ce que que les outils d'ia soient développés pour bénéficier à tous (ou du moins à la plupart) des humains, plutôt qu'à une minorité privilégiée ; qu'ils doivent créer des solutions aux besoins humains réels et améliorer la qualité de vie et l'expérience pour autant d'humains que possible ; et que les dommages réels ou potentiels de ces outils soient reconnus, évalués par rapport à leurs avantages, et atténués chaque fois que possible. Il ne faut pas un cynisme excessif pour reconnaître que beaucoup de ces objectifs ne seront pas atteints dans la pratique ; mais débattre du système de valeurs avec lequel nous souhaitons que ces outils s'alignent est la première étape pour rendre possible la réalisation effective de ces objectifs.

Lorsqu'un certain consensus sera (espérons-le) trouvé sur ces valeurs, puis en coordination avec les actions ci-dessus pour atténuer les pires impacts de l'ia, l'attention devra se porter sur la plus grande source de friction - l'interface entre l'ia et les humains. Pour aller au-delà d'une trêve précaire et instable, nous devons développer des méthodes pour permettre aux individus d'incorporer des outils d'ia dans leurs flux de travail sans risquer de dégrader la valeur de leur travail ou de déposséder les utilisateurs de leur propre humanité.

6. Le spectre des interactions humain-ia

6.1. Une vision à court terme : l'ia comme l'"extrait de vanille" de la production intellectuelle.

Comment devrions-nous conceptualiser l'interface entre l'humanité et les outils d'ia ? Dans l'immédiat, il est encore défendable de considérer ces technologies principalement comme des curiosités, et de nombreux utilisateurs ne savent pas comment raisonnablement les appliquer.

Notre suggestion pour naviguer au milieu de la transition actuelle est d'utiliser une analogie culinaire : l'extrait de vanille, un ingrédient courant dans la plupart des recettes sucrées, célèbre pour son arôme presque universellement attrayant. Ingéré seul, l'extrait de vanille est générale-

ment considéré comme extrêmement désagréable, mais son ajout en petites quantités est largement considéré comme améliorant et rehaussant les autres saveurs du plat, même lorsqu'il ne peut en être différencié. Si on aurait tendance à croire qu'ajouter de l'extrait de vanille améliore un plat, la plupart des gens qui l'ont utilisé comprennent qu'il y a une limite supérieure au-delà de laquelle il ruine complètement le plat [20]. La plupart d'entre nous n'ont pas une idée claire de ce qu'est cette limite supérieure, et trouvent donc plus sage de l'ajouter en très petite quantité.

De même, on pourrait considérer l'utilisation actuelle de l'ia comme un ajout facultatif aux flux de travail cognitifs : il est intéressant d'expérimenter avec modération - une relecture d'un texte composé par un humain via un modèle de langage ia pour des suggestions de grammaire et de formulation, ou une liste de points remise à l'ia pour qu'elle les organise en une structure suggérée. Ces touches légères, comme un petit splash de vanille, amélioreront et enrichiront le caractère de l'œuvre sans la submerger. Le contenu généré par l'ia utilisé comme composant central de tels flux de travail, cependant, ne produira pas de résultats souhaitables, efficaces ou précieux. Avec une telle philosophie (et une citation appropriée de l'utilisation de l'ia), il n'est pas nécessaire de repenser immédiatement les hypothèses fondamentales sur le rôle des humains dans les activités intellectuelles telles que les mathématiques, les sciences ou les arts créatifs.

6.2. Le moyen terme : l'ia dans l'"équipe rouge".

Cependant, à mesure que ces outils augmentent en capacité et sont plus largement adoptés, la possibilité de "se retirer" de ces technologies diminuera. Même si l'on choisit personnellement d'éviter activement l'assistance de l'ia, les collègues, les étudiants et les institutions professionnelles avec lesquels cet individu interagit incorporeront de plus en plus l'ia dans leur propre travail. Actuellement, il y a de sérieuses préoccupations que des domaines entiers du discours académique puissent être noyés par un flot de contenu de faible qualité généré par l'ia. À court terme, cela peut être combattu par des politiques éditoriales strictes interdisant la plupart des formes de contenu généré par l'ia ; mais à mesure que ces outils deviendront plus omniprésents et qu'un réseau d'agents d'ia individualisés deviendra plus courant, une approche plus nuancée deviendra nécessaire.

À moyen terme au moins, il sera encore possible et nécessaire de concevoir des règles et des lignes directrices pour identifier les usages les plus responsables de l'ia et décourager les usages irresponsables, sans changer fondamentalement la nature humaniste de son domaine - en bref, considérer l'assistance de l'ia comme un outil ou un partenaire junior, plutôt que comme un remplacement, pour le travail centré sur l'humain. Dans ce cas, il peut être utile de faire une distinction entre [21] les tâches de "l'équipe bleue" (génération de nouveau contenu et de nouvelles structures) et les tâches de "l'équipe rouge" (vérification, test ou maintenance de ce contenu). L'ia est relativement sûre à utiliser dans une capacité d'"équipe rouge" pour examiner le contenu généré par l'homme à la recherche d'erreurs ou d'améliorations suggérées ; mais avec le manque de fiabilité stochastique et l'absence d'ancrage des outils actuels et à court terme, il n'est pas sûr de leur faire confiance dans une capacité structurelle d'"équipe bleue" qui dépasse la capacité de l'"équipe rouge" (qui pourrait

20. Une expérience de pensée notoire sur `tumblr` [47] a conclu qu'un gâteau contenant 44 % d'extrait de vanille serait immangeable.

21. Cette terminologie s'inspire de la distinction en cybersécurité entre une « équipe bleue » qui défend un système contre les attaquants et une « équipe rouge » qui recherche les faiblesses.

consister en des humains ou des outils de vérification plus automatisés, comme les assistants de preuve formels) à vérifier. Dans cette philosophie, l'accent est mis sur la gestion des risques potentiels de l'utilisation de l'ia tout en capturant encore bon nombre de ses avantages potentiels, plutôt que de repenser radicalement la nature fondamentale du domaine.

6.3. Le long terme : une retraite philosophique est-elle inévitable ?

Mais supposons que l'on regarde vers un avenir plus lointain où les faiblesses actuelles des outils d'ia sont résolues de manière satisfaisante, et où leurs capacités égalent ou dépassent désormais celles [22] des humains experts dans toutes les dimensions pratiques, rendant la philosophie de gestion des risques obsolète. Comment répondrons-nous alors aux questions philosophiques complexes soulevées par la nature transformative de ces technologies avancées ?

Une option est simplement de se retirer dans des cadres purement techniques dans lesquels ces questions ne sont plus opérationnelles. En mathématiques, nous avons le point de vue "formaliste", où le seul objectif est de manipuler des symboles mathématiques selon des règles précises. Dans les sciences, la position philosophique pragmatique "ferme-la et calcule" joue un rôle similaire ; et dans les arts créatifs, on peut travailler comme artisan plutôt que comme artiste, créant des œuvres qui satisfont aux paramètres fournis par un client externe, sans porter de jugement sur la valeur du produit. Dans chacun de ces cas, aucune distinction n'est nécessaire entre le travail généré par l'homme ou par l'ia, tant que les spécifications techniques de la tâche sont remplies.

Mais si la technique est certainement une composante essentielle de chacune de ces disciplines, elle ne capture pas l'expérience complète de la manière dont les mathématiques, les sciences et les arts sont pratiqués dans la réalité, et fournit peu de conseils sur des questions pratiques telles que la manière de motiver la prochaine génération d'étudiants, ou les directions de recherche motivées par la curiosité à poursuivre. On pourrait donc se retirer dans une position radicalement différente, où l'on attribue un statut spécial ineffable à l'intellect humain ou à la créativité humaine, distinguant de manière permanente toute activité exerçant ces traits humains à un niveau fondamental de toute réplication artificielle de cette activité, quelle que soit la précision avec laquelle cette dernière pourrait répliquer ou surpasser la première au niveau technique. Dans ce cadre, l'intelligence artificielle sera à jamais un "vrai Écossais" (No True Scotsman) : manquant de véritable "âme" ou de "compréhension". Avec la longue familiarité avec notre propre espèce, nous sommes habitués à ce que les humains soient peu fiables, "pointus" dans leurs capacités, et parfois réussissent à accomplir une tâche par association de mots aléatoires et mémorisation par cœur ; mais lorsque les outils d'ia présentent des comportements similaires, on peut être enclin à les juger bien plus sévèrement, par exemple en attribuant ces défaillances à leur nature inhérente de "perroquets stochastiques". Mais peut-être que cette position ne fait que nier une vérité inconfortable : une partie de nos capacités humaines vantées n'est en réalité pas beaucoup plus sophistiquée dans sa nature que les algorithmes d'ia que nous avons maintenant conçus pour les imiter. Et à mesure que les performances de l'ia continueront de progresser, un tel point de vue "chauvin humain" risque de dégénérer en une philosophie du "dieu des lacunes" de plus en plus intenable, dans laquelle une liste sans cesse réduite

22. Ce scénario est parfois désigné sous le terme d'"intelligence artificielle générale", bien qu'il n'existe pas de consensus sur la définition précise de ce terme.

de qualités est vantée comme des indicateurs de la réalisation humaine essentielle que l'ia n'est pas encore capable de reproduire.

Une troisième option, particulièrement favorisée par certains enthousiastes de ces technologies, est de soutenir que toutes les capacités cognitives humaines seront bientôt complètement supplantées par leurs équivalents en ia, rendant les discussions philosophiques sur la valeur des contributions et des préoccupations humaines pour les mathématiques, les sciences et les arts de plus en plus sans objet. Dans les versions les plus extrêmes de cette position, l'exercice même de l'intellect humain est considéré comme une activité indésirable et fastidieuse, qui devrait être remplacée par l'automatisation aussi rapidement que possible, afin de libérer du temps et de l'espace mental pour des activités plus hédonistes ou de loisir. Évidemment, une mise en œuvre de cette philosophie comporterait de nombreux risques, comme la dégradation des capacités humaines au point où notre espèce deviendrait collectivement incapable de surveiller, contrôler, ou même comprendre les actions de ces ia de plus en plus puissantes auxquelles nous aurons délégué notre civilisation [23].

Il semble cependant exister un certain terrain d'entente philosophique entre ces extrêmes "hommes de paille", qui peut fournir une perspective utile pour les paradigmes émergents de coopération et de coexistence complémentaire entre les humains et les agents d'ia. Un précédent peut être vu dans le monde des échecs, qui était autrefois considéré comme un exercice quintessentiel de l'intellect humain pur. Cela fait maintenant plusieurs décennies qu'aucun grand maître humain n'a pu battre un moteur d'échecs. Néanmoins, les échecs restent une activité humaine populaire et florissante, les joueurs d'échecs incorporant les moteurs dans leur entraînement, les utilisant pour revisiter d'anciennes théories et en explorer de nouvelles, sonder les exploits et les faiblesses des joueurs d'échecs ia autrement invincibles, ou introduire de manière créative de nouvelles formes de compétition impliquant différents niveaux d'assistance de l'ia. Les questions philosophiques sur ce qu'est réellement le jeu d'échecs, et quelle est la valeur à y jouer, continuent de mériter d'être posées ; et les réponses actuellement acceptées ne ressemblent étroitement à aucune des trois positions extrêmes esquissées ci-dessus.

6.4. Une vision copernicienne.

Une possibilité est d'adopter un analogue cognitif de la révolution copernicienne en astronomie. Dans l'Antiquité, les modèles dominants de la cosmologie (dans la mesure où l'univers était considéré en termes mécanistes) étaient géocentriques, où la Terre avait un statut ontologique privilégié en tant que centre immobile de l'univers, fondamentalement distincte des cieux au-dessus ou des enfers en dessous. Cependant, de multiples avancées en astronomie et en physique ont démantelé cette vision, montrant successivement au fil des siècles que la Terre était en fait en mouvement sur son axe et en orbite autour du Soleil, le Soleil lui-même orbitant autour du centre de notre galaxie, qui à son tour fait partie d'un univers en expansion dépourvu de toute notion de centre spatial. En effet, il est devenu extrêmement fructueux d'adopter un point de vue philosophique complètement opposé, connu aujourd'hui sous le nom de principe copernicien : que la Terre n'est qu'une planète parmi d'innombrables autres dans l'univers, ne recevant aucun traitement préférentiel de la part

23. Pour une vision de ce à quoi ressemblerait ce cadre en pratique, nous suggérons le film de science-fiction Wall-E [48].

des lois fondamentales de la nature.

À première vue, cette vision semble assez menaçante pour l'attachement émotionnel de l'humanité à notre planète natale, mais au fond, il n'y a pas de contradiction fondamentale entre le désintérêt de l'univers pour la planète Terre et notre propre investissement fort en elle ; nous pouvons tout à fait justifier de prioriser les problèmes spécifiques à la planète Terre plutôt que ceux d'autres planètes, tout en acceptant simultanément que ces autres planètes existent et seraient d'une importance comparable pour leurs propres habitants. Des révolutions similaires peuvent être observées dans le développement historique d'autres sciences, par exemple dans la révolution darwinienne détrônant le statut unique des humains parmi d'autres espèces en constante évolution, ou le détrônement du rôle privilégié de la géométrie euclidienne en tant que source de vérité synthétique *a priori* en mathématiques.

Jusqu'à récemment, notre espèce a de même embrassé un analogue intellectuel du modèle géocentrique, dans lequel l'intelligence humaine se tenait au centre de l'univers cognitif, lui conférant ainsi un statut philosophique spécial. Mais maintenant, nous découvrons (ou créons) d'autres "planètes" d'intelligence comparables à bien des égards à la nôtre, tout en étant simultanément assez distinctes à bien des égards. Au lieu de nier l'existence ou l'importance de ces planètes, ou de se disputer pour savoir laquelle mérite d'être le "centre", on peut plutôt accepter que les intelligences humaines et artificielles existent dans la même catégorie ontologique, bien qu'avec de nombreuses différences et complémentarités distinctives. Bien que nos intérêts et attachements restent largement liés à la sphère intellectuelle humaine, sa relation avec d'autres formes d'intelligence peut être explorée, à la fois à des fins pratiques d'atteinte plus efficace de divers objectifs réels, ainsi que pour des raisons plus philosophiques, comme l'obtention d'une perspective externe sur la cognition humaine qui était auparavant difficile à atteindre.

7. Conclusion

La libération non structurée, chaotique et généralisée de la technologie de l'ia dans le monde a déjà considérablement déplacé les sphères sociales, intellectuelles et économiques, d'une manière aussi alarmante que bénéfique. Bien qu'un effort collectif de l'humanité soit sans aucun doute nécessaire, que ce soit par la réglementation, la pression du marché ou par une force encore à définir ; nous n'avons décidément pas encore atteint un point de bascule dont nous ne pourrions nous extirper du coût économique et social élevé de ces nouvelles technologies. Les approches d'intégration de l'ia dans le domaine des mathématiques ont tout aussi rapidement démontré les avantages prometteurs que l'ia peut apporter à la recherche académique, au progrès scientifique et à l'humanité en général. La nature largement objective et vérifiable de la recherche mathématique présente une opportunité unique d'expérimenter ces nouvelles technologies et d'étudier les impacts qui en résultent d'une manière qui ne présente pas de risque éthique ou existentiel pour l'individu ou la société dans son ensemble. À partir de l'application de l'ia aux mathématiques, nous sommes en mesure d'explorer les questions philosophiques et morales pressantes d'une utilisation mondiale plus large de l'ia. De plus, nous pouvons extrapoler des voies potentielles pour soulager les tensions à l'interface ia/humain et suggérer de nouveaux paradigmes de pensée coopérative ia/humain qui respectent les qualités uniques et précieuses que chaque modalité apporte à la table métaphorique. Et si nous ne

pouvons plus jamais enfermer le génie dans la bouteille, nous sommes optimistes sur le fait que, à mesure que nos compréhensions et nos actions progresseront, rapidement, nous pourrons dissiper la fumée et nous tourner vers un avenir radieux, même s'il est aujourd'hui quelque peu incertain.

Remerciements

Nous remercions Silvia de Toffoli pour ses commentaires et références utiles.

Références

[1] J. Jumper, R. Evans, A. Pritzel, et al., "Highly accurate protein structure prediction with `alphafold`," *Nature*, vol. 596, pp. 583-589, Aug. 2021.

[2] S. Marche, "The College Essay Is Dead," Dec. 2022.

[3] D. Anguiano and L. Beckett, "How Hollywood writers triumphed over ai - and why it matters," *The Guardian*, Oct. 2023.

[4] E. Oh, W. Kearns, M. Laine, G. Demiris, and H. J. Thompson, "Perceptions of and Experiences with Consumer Sleep Technologies That Use Artificial Intelligence," *Sensors*, vol. 22, p. 3621, Jan. 2022.

[5] F. Swain, "Is it right to use Nazi research if it can save lives ?" https://www.bbc.com/future/article/20190723-the-ethics-of-using-nazi-science.

[6] A. Tarkowski, "Open source and the democratization of ai," in *Artificial Intelligence and the Challenge for Global Governance : Nine Essays on Achieving Responsible ai* (A. Krasodomski, ed.), pp. 30-36, Royal Institute of International Affairs, June 2024.

[7] J. Chun and K. Elkins, "The Crisis of Artificial Intelligence : A New Digital Humanities Curriculum for Human-Centred ai," *International Journal of Humanities and Arts Computing*, vol. 17, pp. 147-167, Oct. 2023.

[8] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, pp. 417-424, Sept. 1980.

[9] Q. Mei, Y. Xie, W. Yuan, and M. O. Jackson, "A Turing test of whether ai chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, vol. 121, p. e2313925121, Feb. 2024.

[10] H. Chen, S. R. Grimm, O. Russakovsky, and T. Lombrozo, "Machine understanding," Unpublished preprint.

[11] M. Mashaal, *Bourbaki : A Secret Society of Mathematicians*. Providence, RI : American Mathematical Society, 2006.

[12] *The mathematical experience*. Boston : Birkhäuser, 1981.

[13] R. Wagner, "Mathematical consensus : A research program," *Axiomathes*, vol. 32, pp. 1185-1204, Dec. 2022.

[14] J. Baez, "The Inconsistency of Arithmetic." https://golem.ph.utexas.edu/category/2011/09/the_inconsistency_of_arithmeti.html.

[15] S. Aaronson, "Ten Signs a Claimed Mathematical Breakthrough is Wrong," Jan. 2008.

[16] S. DeDeo, "AlephZero and mathematical experience," *Bulletin of the American Mathematical Society*, vol. 61, pp. 375-386, July 2024.

[17] T. Tao, "What is good mathematics ?" *Bulletin of the American Mathematical Society*, vol. 44, no. 4, pp. 623-634, 2007.

[18] W. P. Thurston, "On Proof and Progress in Mathematics," in *18 Unconventional Essays on the Nature of Mathematics* (R. Hersh, ed.), pp. 37-55, New York, NY : Springer, 2006.

[19] S. de Toffoli and F. Tanswell, "The technological turn in mathematics," *Blackwell Companion to the Philosophy of Mathematics*, 2025.

[20] Y. Wu, A. Q. Jiang, W. Li, M. Rabe, C. Staats, M. Jamnik, and C. Szegedy, "Autoformalization with Large Language Models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32353-32368, Dec. 2022.

[21] F. Tanswell, "Can Mathematics Be Hacked ? Infrastructure, Artificial Intelligence, and the..." June 2025.

[22] K. I. Appel and W. Haken, *Every Planar Map Is Four Colorable*, vol. 98. American Mathematical Soc., 1989.

[23] T. C. Hales, "A Proof of the Kepler Conjecture," *Annals of Mathematics*, vol. 162, no. 3, pp. 1065-1185, 2005.

[24] A. Venkatesh, "Some thoughts on automation and mathematical research," *Bulletin of the American Mathematical Society*, vol. 61, pp. 203-210, Feb. 2024.

[25] S. DeDeo, "Hard Proofs and Good Reasons," Oct. 2024.

[26] J. Avigad, "Is mathematics obsolete ?," 2025.

[27] "Special issue on ai and mathematics, Part I," *Bulletin of the American Mathematical Society*, vol. 61, pp. 199-372, Apr. 2024.

[28] "Special issue on ai and mathematics, Part II," *Bulletin of the American Mathematical Society*, vol. 61, pp. 373-524, July 2024.

[29] H. Macbeth, "Algorithm and abstraction in formal mathematics," May 2024.

[30] J. Stillwell, *Reverse Mathematics : Proofs from the inside Out*. Princeton, New Jersey : Princeton University Press, 2018.

[31] A. Jaffe and F. Quinn, "Theoretical mathematics" : Toward a cultural synthesis of mathematics and theoretical physics," *Bulletin of the American Mathematical Society*, vol. 29, no. 1, pp. 1-13, 1993.

[32] M. Kim, "Thinking and explaining," MathOverflow. https://mathoverflow.net/q/38694 (version : 2024-01-05).

[33] T. Hey, "The Fourth Paradigm - Data-Intensive Scientific Discovery," in *E-Science and Information Management* (S. Kurbanoglu, U. Al, P. L. Erdogan, Y. Tonta, and N. Uçak, eds.), vol. 317, pp. 1-1, Berlin, Heidelberg : Springer Berlin Heidelberg, 2012.

[34] "ai achieves silver-medal standard solving International Mathematical Olympiad problems." https://deepmind.google/blog/ai-solves-imo-problems-at-silver-medal-level/, May 2024.

[35] C. J. Rittberg, "Justified epistemic exclusions in mathematics," *Philosophia Mathematica*, vol. 31, pp. 330-359, 04 2023.

[36] S. De Toffoli and F. S. Tanswell, "Trust in mathematics," *Philosophia Mathematica*, pp. 1-25, 2025. Published online ahead of print.

[37] "Copyright and Fair Use | Office of the General Counsel." https://ogc.harvard.edu/pages/copyright-and-fair-use.

[38] A. Weir, "Chapter 11," in *Project Hail Mary*, Penguin Books (Series), pp. 191-194, London : Penguin Books, 2022.

[39] "Best practices for incorporating ai etc. in papers." https://ai-math.zulipchat.com/.

[40] S. Bubeck, C. Coester, R. Eldan, T. Gowers, Y. T. Lee, A. Lupsasca, M. Sawhney, R. Scherrer, M. Sellke, B. K. Spears, D. Unutmaz, K. Weil, S. Yin, and N. Zhivotovskiy, "Early science acceleration experiments with GPT-5," Nov. 2025.

[41] T. F. Bloom, "Erdős Problems." https://www.erdosproblems.com/.

[42] J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi, "The ai gambit : Leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations," *ai and society*, vol. 38, pp. 283-307, Feb. 2023.

[43] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "ai models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755-759, July 2024.

[44] "Supercharging Research : Harnessing Artificial Intelligence to Meet Global Challenges | PCAST," tech. rep., President's Council of Advisors on Science and Technology, June 2024.

[45] E. Jones, "A 'CERN for ai' - what might an international ai research organization address ?" in *Artificial Intelligence and the Challenge for Global Governance : Nine Essays on Achieving Responsible ai* (A. Krasodomski, ed.), pp. 20-29, Royal Institute of International Affairs, June 2024.

[46] M. Mantegna, "An ethics framework for the ai-generated future," in *Artificial Intelligence and the Challenge for Global Governance : Nine Essays on Achieving Responsible ai* (A. Krasodomski, ed.), pp. 47-57, Royal Institute of International Affairs, June 2024.

[47] "Vanilla Extract." https://knowyourme.com/memes/vanilla-extract, Feb. 2023.

[48] "WALL-E," 2008.