IA Geoffrey Hinton, l'un des pionniers de l'Intelligence Artificielle

Geoffrey Everest Hinton (né le 6 décembre 1947) est un informaticien , un scientifique cognitif et un psychologue cognitif britannico-canadien connu pour ses travaux sur les réseaux neuronaux artificiels , qui lui ont valu le titre de « parrain de l'IA ».

Geoffrey Hinton

Sana Vidéo de 45 mn (mai 2024) ♦♦

Geoffrey Hinton reveals the surprising truth about AI's limits and potential

In this wide-ranging interview, the Godfather of AI Geoffrey Hinton reflects on the journey from the early days of neural networks to today's breakthroughs in artificial intelligence, sharing unique insights from decades of pioneering research.

Dans cette interview exhaustive, Geoffrey Hinton, le parrain de l'IA, revient sur le parcours qui a mené des débuts des réseaux neuronaux aux percées actuelles en intelligence artificielle, partageant des perspectives uniques issues de décennies de recherche pionnière.

Chapitres et analyse

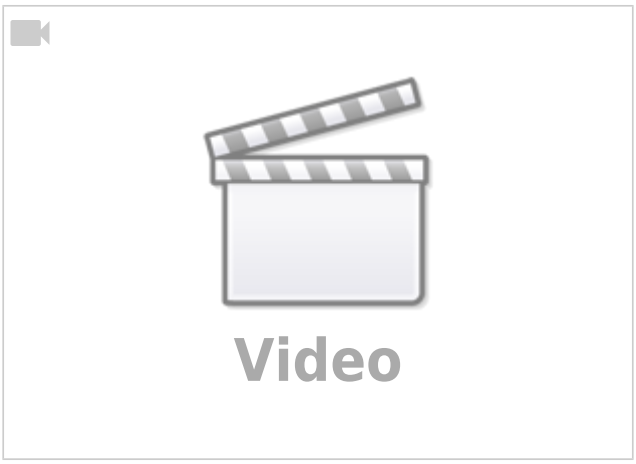
IASEAI Vidéo de 18mn (2025)

What Is Understanding? – Geoffrey Hinton | IASEAI 2025 ♦♦♦

In this groundbreaking plenary from IASEAI '25, Geoffrey Hinton—a Turing Awardee and Nobel Laureate—delves into the concept of "understanding" within machine learning. Drawing on deep neural networks and cognitive science, Hinton investigates whether current models truly grasp meaning or rely on statistical correlations—and why that distinction matters for future AI safety and alignment.

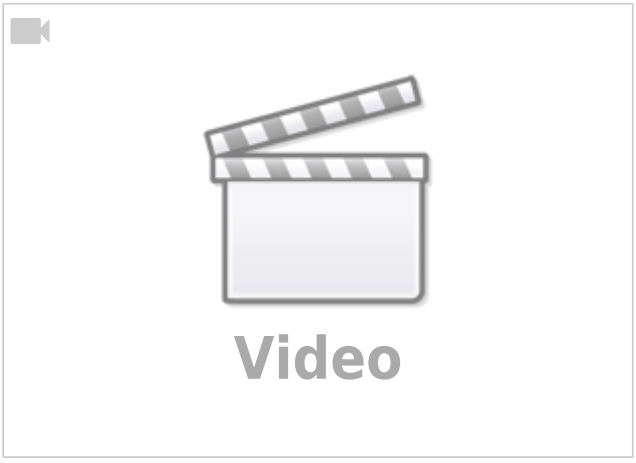
Dans cette conférence plénière novatrice de l'IASEAI 2025, Geoffrey Hinton, lauréat du prix Turing et

du prix Nobel, explore le concept de « compréhension » dans l'apprentissage automatique. S'appuyant sur les réseaux neuronaux profonds et les sciences cognitives, il examine si les modèles actuels saisissent véritablement le sens ou s'ils se contentent de corrélations statistiques, et pourquoi cette distinction est cruciale pour la sécurité et l'alignement futurs de l'IA.



Lien Youtube

“Godfather of AI” Geoffrey Hinton: The 60 Minutes Interview



Lien Youtube

Risque existentiel posé par l'intelligence artificielle

Rage Against the Machine article dans Torento Life ♦

Geoffrey Hinton spent half a century developing artificial intelligence. Now, he worries that his life’s work could spell the end of humanity. Inside his mission to warn the world

By Luc Rinaldi| Portrait by Markian Lozowchuk | November 16, 2023

Rage Against the Machine

Traduction google

En 2023, l'intelligence artificielle a enfin tenu toutes ses promesses. En novembre dernier, le laboratoire de recherche américain OpenAI a lancé ChatGPT, le chatbot désormais omniprésent. Capable de résumer des romans en quelques secondes et d'écrire du code informatique, son potentiel pour générer des scénarios a même contribué à la grève des scénaristes hollywoodiens. En deux mois seulement, il comptait 100 millions d'utilisateurs, devenant ainsi l'application à la croissance la plus rapide de tous les temps, et Microsoft a investi 10 milliards de dollars dans OpenAI pour pérenniser ce succès. Après des décennies de faux départs, l'IA était enfin lancée.

Il y avait cependant une personne qui ne sabrait pas le champagne : Geoffrey Hinton , professeur d'informatique à l'Université de Toronto, plus connu comme le parrain de l'IA. Sur le papier, ChatGPT aurait dû ravir Hinton : il avait consacré toute sa carrière à perfectionner les réseaux neuronaux, l'architecture qui sous-tend GPT, et maintenant, ils fonctionnaient mieux que jamais. Lorsqu'il soumettait des blagues au chatbot, celui-ci pouvait en expliquer le sens. S'il lui proposait des énigmes, le chatbot pouvait les résoudre. « C'est bien plus de raisonnement que ce que nous pensions possible il y a quelques années », dit-il. Il lui semblait que, pour la première fois, les machines réussissaient le test de Turing , le test de référence où les ordinateurs démontrent une intelligence indiscernable de celle d'un humain. Il ne faudrait pas longtemps – peut-être cinq à vingt ans, pensait-il – pour que l'IA devienne totalement plus intelligente que les humains.

Cette prédiction recèle des implications terrifiantes. L'humanité domine la Terre depuis des millénaires précisément parce qu'elle est l'espèce la plus intelligente. Que signifierait l'émergence d'une forme d'intelligence supérieure ? Certes, l'IA pourrait guérir des maladies, atténuer le changement climatique et améliorer la vie sur Terre de manières encore inimaginables – si nous parvenons à la contrôler. Dans le cas contraire ? Hinton craint le pire : que les machines prennent le pouvoir. « Je ne pense pas que nous ayons la moindre chance de garder le contrôle si elles le veulent », affirme Hinton. « Ce sera sans espoir. »

Hinton se demandait quoi faire. Convaincu que l'IA pourrait bien mener l'humanité au bord du précipice, il ne pouvait se résoudre à poursuivre ses travaux. Aussi, le 2 mai, il fit la une du New York Times pour annoncer sa démission de Google et alerter le monde sur la menace existentielle que représente l'IA.

Hinton n'était pas le premier à prophétiser une apocalypse de l'IA. Elon Musk , par exemple, a passé des années à rabâcher l'imminence de la singularité, ce point où les humains perdraient irrémédiablement le contrôle de l'IA – mais Elon dit souvent des choses farfelues. Parmi les experts en IA, rares sont ceux qui ont pris au sérieux l'idée que les machines deviendraient extrêmement dangereuses dans un avenir proche.

Hinton a changé la donne. Après tout, nul n'est plus grand expert en IA. Britannique de naissance et Canadien d'adoption, il a été impliqué, directement ou indirectement par le biais de ses étudiants et collègues, dans presque toutes les grandes avancées de l'apprentissage profond, notamment le développement d'outils d'IA générative comme DALL-E. Lorsqu'il a pris la parole, le monde entier l'a écouté. Jeff Clune, professeur associé à l'Université de Colombie-Britannique et conseiller principal en recherche chez DeepMind, le laboratoire de recherche en IA de Google, m'a confié que l'avertissement de Hinton avait été un véritable coup de tonnerre, ouvrant les yeux des scientifiques, des organismes de réglementation et du grand public. « Il y a des gens des deux côtés de ce débat. Ce qui est rare, c'est de voir quelqu'un changer de camp, et cela attire l'attention », explique-t-il. « Quand cette personne est la plus influente du domaine et, à bien des égards, son père fondateur, il

est impossible de l'ignorer. »

Hinton espérait qu'en tirant la sonnette d'alarme, il inciterait les décideurs politiques à accélérer les efforts pour prévenir un cataclysme lié à l'IA, mais même lui n'avait pas anticipé le raz-de-marée d'attention que son annonce susciterait. Justin Trudeau l'invita à dîner à la gare Richmond de Toronto pour discuter des mesures que le Canada devrait prendre. Leur conversation dura finalement deux heures et demie. S'appuyant en partie sur cette rencontre, le gouvernement Trudeau a mis en place un code de conduite pour les entreprises technologiques , qui encourage (sans toutefois l'imposer pour l'instant) ses signataires – parmi lesquels BlackBerry, Telus et l'Institut Vector – à mettre en œuvre des stratégies robustes de gestion des risques, à publier ouvertement des informations sur leurs systèmes d'IA et à maintenir une supervision humaine. Le gouvernement fédéral espère rendre ces règles obligatoires en adoptant le projet de loi C-27, qui contient la Loi sur l'intelligence artificielle et les données, en 2024.

Début mai, Hinton a reçu un appel de Margrethe Vestager, vice-présidente exécutive de la Commission européenne, qui a depuis intégré plusieurs mesures de protection de l'IA dans son Règlement général sur la protection des données (RGPD). Le gouvernement britannique a convoqué Hinton à Downing Street où, sous un imposant portrait de Margaret Thatcher, il a déclaré à une douzaine de conseillers du Premier ministre Rishi Sunak que, compte tenu du nombre d'emplois que l'IA allait supprimer, ils devraient envisager la mise en place d'un revenu universel. « Surtout, ne dites pas à Sunak que ça s'appelle du socialisme », a-t-il ajouté. À peu près au même moment, Hinton s'est entretenu avec Bernie Sanders au sujet des sanctions potentielles pour la création de désinformation alimentée par l'IA. Il a ensuite discuté avec le sénateur Jon Ossoff, le cabinet de Chuck Schumer et, en juillet, la Maison Blanche, qui l'a informé – « avant même que le Congrès ne soit au courant », précise-t-il – que plusieurs géants américains, dont Google, Amazon, Meta, Microsoft et OpenAI, avaient signé une nouvelle série d'engagements volontaires en matière de sécurité de l'IA. Mais il a décliné l'invitation à comparaître devant une commission de la Chambre des représentants présidée par Jim Jordan, cofondateur du Freedom Caucus. « Les Républicains raffolent des fausses informations », affirme-t-il. « Ils s'en nourrissent. »

Au-delà des instances bureaucratiques, Hinton a provoqué un véritable séisme dans les médias, les affaires et la culture populaire. Le comité éditorial du New York Times l'a invité à participer à l'élaboration de sa politique relative aux contenus et images générés par l'IA – des directives que les rédactions du monde entier ne manqueront pas d'imiter. Il a également reçu un appel de Musk, qui, tout sourire, a acquiescé, convaincu que les machines allaient bientôt prendre le contrôle de l'humanité. « Selon lui, on nous garderait en vie par curiosité, et je pensais que c'était un fil ténu pour notre propre existence », explique Hinton. « Il n'arrêtait pas de parler. J'ai fini par lui dire que j'avais une autre réunion. »

Hinton est désormais une sorte de célébrité alarmiste, exprimant ses craintes dans un flot incessant d'interviews, de podcasts, de conférences et de tables rondes. Il est apparu sur CNN, PBS, CBC, BBC et dans l'émission 60 Minutes . Un journaliste du New Yorker l'a même accompagné dans sa maison de campagne cet été. Snoop Dogg, s'exprimant lors d'une conférence à Beverly Hills, a rapporté : « J'ai entendu le vieux qui a créé l'IA dire : “Ce n'est pas sûr, parce que les IA ont leur propre conscience, et ces enfoirés vont commencer à faire n'importe quoi.” » Ce à quoi Hinton a répondu plus tard, avec son humour britannique pince-sans-rire : « Ils n'avaient probablement pas de mère. »

De toute évidence, Hinton s'amuse. Mais certains de ses amis et collègues – notamment ceux qui ne partagent pas ses craintes apocalyptiques – craignent qu'il ne ternisse son héritage. Avant que Hinton ne prenne la parole publiquement, Aaron Brindle, son attaché de presse de longue date chez Google, l'a conseillé avec tact : en se rangeant du côté des Cassandres, il risquait d'éclipser tout le reste de

son œuvre. « Je pense que son rôle dans l'IA est bien plus important que cela », affirme Brindle, aujourd'hui associé chez Radical Ventures, une société de capital-risque spécialisée dans l'IA.

Hinton restait imperturbable. « Je me fiche de ma réputation », dit-il. « Le mieux qu'on puisse faire avec une bonne réputation, c'est de la dilapider, car on ne l'emporte pas avec soi dans la tombe. »

Ces derniers mois, Hinton a reçu plus d'un millier de demandes d'entrevue. Lorsqu'il a accepté de s'entretenir avec Toronto Life , il a exigé que le journaliste soit titulaire d'un diplôme en sciences, technologies, ingénierie et mathématiques (STEM) afin de pouvoir approfondir les fondements techniques de l'intelligence artificielle. Arrivé chez Hinton, dans sa charmante maison de briques du quartier Annex, en septembre dernier, j'avoue, un peu gêné, mon manque de qualifications. « Pour la petite histoire », lui dis-je, « j'étais un as en mathématiques au secondaire. » « Vous savez donc ce qu'est un polynôme », conclut-il, à tort. Pour la première fois de ma vie, je regrette d'avoir abandonné le calcul différentiel et intégral en terminale.

La maison est calme. Les deux enfants adultes de Hinton, qui vivent avec lui, sont absents. Il les a adoptés en Amérique latine dans les années 1990 avec sa première épouse, Ros, décédée d'un cancer des ovaires alors qu'ils étaient en bas âge. Sa seconde épouse, Jackie, est décédée d'un cancer du pancréas il y a cinq ans. Hinton a une nouvelle compagne, Rosemary Gartner, professeure de criminologie à l'Université de Toronto. Mais, le jour de ma visite, ses deux chats sont les seuls autres êtres vivants présents.

Nous nous installons à une longue table en bois dans la salle à manger – ou plutôt, je m’assieds. Hinton s’est blessé au dos à 19 ans en déplaçant un radiateur d’appoint pour sa mère et a cessé de s’asseoir complètement en 2005, la douleur étant devenue insupportable. Ces derniers temps, il parvient à rester assis par tranches de 15 minutes ; ainsi, tout au long de nos deux heures de conversation, il alterne entre des allers-retours autour de la table et des moments où il s’assoit sur une boîte à chaussures posée sur une chaise.

Avant ma visite, j'ai discuté avec plusieurs collègues de Hinton, qui l'ont décrit tour à tour comme un « penseur profond », un être « d'un autre monde » et même « peut-être un extraterrestre ». Juna Kollmeier, professeure d'astronomie à l'Université de Toronto, m'a confié : « S'il existe une vie intelligente dans l'univers, c'est bien lui. » L'intelligence coule dans ses veines. Son arbre généalogique compte un nombre impressionnant de scientifiques d'une influence considérable, parmi lesquels le créateur de la logique booléenne, l'inventeur des structures de jeux et celui qui a donné son nom au mont Everest. Par nature, par éducation, ou les deux, Hinton possède l'instinct du scientifique pour comprendre le fonctionnement du monde – une curiosité polymathe qui s'applique aussi bien à l'informatique qu'à la menuiserie. Moins d'une demi-heure après mon arrivée, il avait déjà disserté sur la psychologie pédiatrique, l'acide pyruvique, la dentisterie, le Watergate, la faillibilité de la mémoire humaine et, bien sûr, l'intelligence artificielle .

Je demande à Hinton comment il en est venu à penser que l'IA représente une menace existentielle pour l'humanité. Tout a commencé, explique-t-il, lorsqu'il a entrepris de résoudre l'un des problèmes les plus épineux de l'IA : sa consommation énergétique astronomique. Selon une estimation, ChatGPT — qui, comme la plupart des modèles d'apprentissage automatique, repose sur une multitude de serveurs informatiques très gourmands en énergie — consomme un gigawattheure d'électricité par jour , soit suffisamment pour alimenter plus de 30 000 foyers. À titre de comparaison, le cerveau humain consomme environ 12 watts, soit moins qu'une ampoule classique. Cette différence s'explique par le fait que les machines apprennent différemment de nous. La plupart des IA modernes utilisent un algorithme appelé rétropropagation, qui consiste à faire passer des données — les pixels d'une image, par exemple — à travers un réseau de neurones artificiels et à ajuster les connexions entre ces neurones jusqu'à ce que la machine puisse reconnaître des caractéristiques comme les

formes et les couleurs et dire, par exemple : « C'est un chat. » Ces calculs activent des milliards de minuscules transistors en silicium, qui génèrent d'énormes quantités de chaleur et nécessitent, par conséquent, des systèmes de refroidissement énergivores.

Hinton a tenté de créer une version d'IA moins énergivore, plus proche du fonctionnement du cerveau humain. Mais à mesure que son IA se rapprochait du fonctionnement humain, elle régressait. Certes, elle consommait moins d'énergie, mais elle était aussi moins performante. Les grands modèles d'IA peuvent traiter d'énormes quantités de données – livres, pages web, extraits audio et vidéos – bien plus rapidement que les humains, et transmettre leurs connaissances à des dizaines de milliers d'ordinateurs, une sorte d'« intelligence collective », selon les termes de Hinton. L'IA utilisant l'algorithme de Hinton, quant à elle, transmettait ses connaissances lentement, à la manière d'un professeur enseignant à un élève.

Cela a conduit Hinton à une constatation surprenante : l'intelligence artificielle fonctionnait bien mieux et plus vite que l'intelligence biologique. Pour lui, c'était un changement de paradigme. Il avait toujours pensé que, dans un avenir prévisible, le cerveau humain serait supérieur à l'IA. « C'est ce que je croyais l'année dernière, et l'année d'avant, et pendant les 48 années précédentes », dit-il. « J'ai soudain réalisé que je me trompais. »

Pourtant, je ne comprenais pas bien comment l'IA que nous connaissons aujourd'hui — tous ces chatbots qui produisent des dissertations universitaires médiocres — pourrait évoluer en superpuissances capables d'exterminer des espèces. « On pourrait penser que nous les ferons en sorte qu'elles ne souhaitent jamais prendre le pouvoir », explique Hinton. Mais les humains devront nécessairement définir des objectifs pour l'IA. Et pour lui permettre d'atteindre ces objectifs efficacement, ajoute-t-il, nous devrons lui donner la capacité de prendre des décisions sans intervention humaine. En réalité, c'est déjà le cas : certaines entreprises technologiques autorisent l'IA à acquérir de nouveaux espaces serveur sans validation humaine. Et il est probable que les IA seront un jour capables de modifier leur propre code, ce qui leur donnera encore plus d'autonomie. « C'est comme pour les soldats », explique Hinton. « Un général n'a pas besoin de dire : "Pointez votre arme ici, et quand je vous le dirai, appuyez sur la détente." Le général dit simplement : "Tuez l'ennemi", et les soldats obéissent. » De même, soutient-il, les machines dotées d'IA développeront de manière autonome leurs propres sous-objectifs. « Le problème, c'est que tout objectif a naturellement pour but d'obtenir plus de contrôle. Si on leur demande d'être efficaces, ils vont vouloir plus de contrôle. Et c'est le début d'une pente glissante. »

Je me demandais à quoi ressemblerait le bas de cette pente. HAL-9000 éliminant des astronautes pour accomplir sa mission ? Des cyborgs de Skynet massacrant des humains sans défense ? Pour en savoir plus sur la manière dont nos maîtres IA pourraient prendre le pouvoir, j'ai discuté avec David Duvenaud, collègue de Hinton et professeur agrégé à l'Université de Toronto, spécialiste de la sécurité de l'IA. Duvenaud m'a expliqué qu'une prise de contrôle par l'IA ne signifie pas forcément des machines malveillantes déterminées à asservir les humains. Selon lui, nous allons plutôt nous marginaliser progressivement et imperceptiblement. Nous dépendons déjà des machines pour nous aider à choisir un candidat ou à investir en bourse. Avec les progrès de l'apprentissage automatique, les pays et les entreprises devront soit adopter l'IA, soit risquer de se laisser distancer par leurs concurrents. Bientôt, nous nous tournerons vers les machines pour décider de la gestion d'une entreprise ou de la conduite d'une guerre. « Il sera souvent plus logique de remplacer un humain par une machine », affirme Duvenaud. Et à mesure que nous déléguons de plus en plus de pouvoir de décision à l'IA, « nous serons progressivement écartés des secteurs clés et stratégiques de notre civilisation ». Cela ne signifie pas pour autant l'extinction. Mais, ajoute-t-il, « vivre dans une civilisation où l'on n'apporte quasiment rien à personne est la garantie d'une perte de pouvoir et d'influence à long terme ».

« Le problème, c'est que tout objectif a naturellement pour but d'obtenir plus de contrôle. Si on demande à l'IA d'être efficace, elle voudra forcément plus de contrôle. Et c'est le début d'une pente glissante. » D'accord, mais si l'humanité n'apprécie pas la direction que prend l'IA, ne pourrons-nous pas la désactiver ? Le PDG d'OpenAI, Sam Altman, transporterait soi-disant un interrupteur d'arrêt d'urgence dans un petit sac à dos bleu pour désactiver ChatGPT en cas de problème. Après tout, on parle d'ordinateurs. J'ai posé la question à Hinton, mais il ne semble pas rassuré. Les IA superintelligentes, dit-il, seront capables de nous surpasser et de nous manipuler pour que nous fassions leur volonté. Elles pourraient même se faire passer pour moins intelligentes. « Elles parviendront à convaincre celui qui a l'interrupteur d'arrêt d'urgence de ne pas l'actionner », dit-il. « Imaginez que vous viviez dans un monde d'enfants de deux ans, et que ces enfants soient au pouvoir. Vous trouveriez des moyens de les manipuler. "Donnez-moi le pouvoir. Il y a des bonbons gratuits pour tout le monde !" Et voilà. » Pour détendre l'atmosphère, il ajoute : « Vous pensez que je devrais m'acheter un de ces chapeaux qu'Oppenheimer portait ? »

Hinton se lève pour nous préparer du thé. Tout en mettant la bouilloire en marche, il explique que, durant l'été, plusieurs agences de conférenciers l'ont contacté pour lui proposer des conférences rémunérées. « Je me suis dit que j'allais en tenter une, juste pour voir », dit-il. Il a choisi l'offre la plus lucrative, qui se trouvait à Las Vegas. Je souris en imaginant Hinton, l'universitaire si sérieux, s'éclater dans la Cité du Péché. Je lui demande s'il y est déjà allé. Une seule fois, répond-il, lors d'un périple de 17 700 kilomètres en Greyhound à travers l'Amérique, à l'âge de 17 ans. « J'ai mis 25 cents dans une machine à sous et j'ai gagné un dollar », dit-il. « Et puis j'ai arrêté. »

Deux semaines plus tard, je me suis envolé pour Las Vegas afin d'assister à la conférence Info-Tech Live, animée par Hinton. Organisée par Info-Tech Research Group, une entreprise de London (Ontario) proposant des services de recherche et de conseil exclusifs aux professionnels de l'informatique et aux DSI, cette conférence de trois jours se tenait à l'hôtel-casino Cosmopolitan. Récemment, la firme s'était spécialisée dans l'accompagnement des entreprises dans la mise en œuvre de stratégies d'apprentissage automatique. De ce fait, près de la moitié des ateliers et tables rondes de la conférence étaient consacrés à l'IA. Un intervenant a décrit comment un modèle d'IA avait conçu le plan d'un immeuble de bureaux à Toronto. Un autre a présenté l'utilisation de l'IA pour détecter les appels téléphoniques falsifiés. Lors de la première conférence plénière, le futurologue Ray Kurzweil a vanté devant un auditoire de 1 500 personnes tous les bienfaits que l'IA apportera à l'humanité : une vie plus longue, des démocraties plus fortes, des voitures plus sûres, des journées de travail plus courtes et des revenus plus élevés. « Son développement sera exponentiel et elle permettra de résoudre les problèmes médicaux mille fois plus vite que les techniques conventionnelles », a-t-il déclaré. D'ici les années 2030, prédit-il, les humains pourront connecter leur cerveau au cloud, ce qui nous permettra d'exploiter la puissance de l'IA sans même avoir à appuyer sur une touche.

Alors que Kurzweil avait enflammé la foule avec son discours sur l'IA, la conférence de Hinton l'après-midi même a douché leurs espoirs – et le moindre optimisme. Il est apparu sur scène tel un fantôme (pull noir, pantalon noir, baskets noires) et a fixé le vide d'un air sombre pendant que le présentateur l'introduisait. Après quelques questions sur sa carrière, l'intervieweur l'a interrogé sur la différence entre intelligence biologique et intelligence numérique. « Ce sera une mauvaise nouvelle pour Ray Kurzweil », a répondu Hinton. On peut facilement transplanter un modèle d'apprentissage automatique sur un nouvel ordinateur où il fonctionnera de la même manière, a-t-il expliqué, ce qui rend le savoir numérique immortel. Mais c'est impossible avec le cerveau humain. « Par conséquent, tout ce que Ray sait disparaîtra à sa mort – et il mourra. »

Après avoir passé en revue les dangers de l'IA, Hinton a ajouté ses mises en garde habituelles. « Nous entrons dans une période d'immense incertitude. De nombreuses possibilités dystopiques et déprimantes existent, mais nous ne savons rien avec certitude », a-t-il déclaré. « Nous, et en particulier les hommes blancs d'un certain âge, avons l'habitude de nous considérer comme les

maîtres de la situation. Nous avons du mal à accepter que ces entités puissent être bien plus intelligentes que nous et décider de se passer de nous. Nous devons absolument empêcher cela. » L'intervieweur a insisté : comment éviter ce sort ? Visiblement abattu, Hinton a admis son ignorance. Il n'était même pas certain que ce soit possible. On ne peut pas effacer des décennies de progrès technologique et reléguer l'IA au rang de science-fiction. « Je n'ai pas de solution à ces problèmes », a-t-il affirmé. « J'aimerais que ce soit comme pour le changement climatique, où l'on peut dire : “Arrêtons de brûler du carbone”. Il n'existe pas de recette miracle pour l'IA. »

Avant la fin de l'entretien, l'intervieweur a demandé à Hinton quels secteurs l'IA pourrait bouleverser. « La réponse est assez courte : tous. » Quelques minutes plus tard, il a nuancé sa réponse : la plomberie est épargnée pour l'instant. « Surtout la plomberie dans une vieille maison, car il faut être très ingénieux et agile, et avoir les doigts dans des endroits difficiles d'accès, ce qui n'est pas encore très efficace. » Une foule d'informaticiens inquiets ont ri nerveusement.

Ce soir-là, j'ai rejoint Hinton et une demi-douzaine de cadres d'Info-Tech pour dîner dans un restaurant de tapas. Il se tenait à une extrémité de la table, picorant de la paella et répondant aux questions. Quelqu'un lui a demandé ce qu'il était devenu depuis son départ de Google. « Plomberie », a-t-il répondu sans la moindre ironie, avant de se lancer dans un monologue détaillé de dix minutes, photos à l'appui, sur la façon dont il avait réparé une fuite dans ses toilettes à l'étage. Un peu plus tard, tentant de ramener la conversation au sujet du jour, quelqu'un a demandé à Hinton quelle était, selon lui, l'opportunité la plus passionnante offerte par l'IA. Avec un sourire en coin, il a plaisanté : « Qu'elle nous tue tous. »

La grande question, bien sûr, est : que faire maintenant ? Fin octobre, Hinton a proposé une voie à suivre. Dans une lettre ouverte , avec 23 autres experts internationaux, il a appelé les principaux laboratoires d'IA à consacrer un tiers de leurs budgets de R&D à garantir la sécurité et l'éthique de leurs systèmes. Ils ont également conseillé aux gouvernements, entre autres, de créer un registre des grands systèmes d'IA, d'obliger les entreprises à signaler les cas d'IA présentant un comportement dangereux et de protéger juridiquement les lanceurs d'alerte. Il est trop tôt pour dire si les laboratoires d'IA et les législateurs tiendront compte de ces recommandations. Mais Hinton, à 75 ans, s'est résigné à l'idée qu'il ne sera plus très longtemps à la tête de ce combat. La tâche ingrate de sauver le monde incombera à la prochaine génération.

Quelqu'un a demandé à Hinton quelle était, selon lui, l'opportunité la plus excitante offerte par l'IA. « Qu'elle nous tue tous », a-t-il plaisanté. Le meilleur espoir de l'humanité repose peut-être sur Ilya Sutskever, ancien étudiant de Hinton. En juillet dernier, il a annoncé la création et la codirection d'une nouvelle équipe chez OpenAI, baptisée Superalignment. Cette équipe propose une approche novatrice de l'alignement, ce domaine de recherche visant à empêcher l'IA de dérailler et à garantir qu'elle serve des objectifs éthiques et humains. Il ne s'agit pas d'un projet mené en parallèle d'un projet personnel. Sutskever, l'un des plus grands experts mondiaux en apprentissage profond, est cofondateur et directeur scientifique d'OpenAI. Il consacre un cinquième de la puissance de calcul de l'entreprise à la résolution de ce problème, tandis que son supérieur, Sam Altman, parcourt le monde pour sensibiliser les présidents et les premiers ministres aux risques liés à l'IA.

Lors de mon appel vidéo avec Sutskever en septembre, il semblait avant tout très occupé. Il parlait vite et regardait constamment par-dessus son épaule, comme si plusieurs programmeurs, chargés de sauver le monde, avaient besoin de son attention. « Il y a tellement de défis techniques différents à relever », dit-il. N'ayant, comme je l'ai mentionné précédemment, aucun diplôme en informatique, je lui ai demandé de me faire un résumé, en termes simples, de ce qu'il espérait que l'équipe de Superalignment accomplirait. « C'est comme si l'on voulait imprégner la superintelligence d'une force, d'une précision et d'une longévité incroyables » — cette empreinte étant le désir de servir les

humains plutôt que de leur prendre le contrôle.

En parlant avec Sutskever, j'ai presque l'impression d'entendre le tic-tac de l'horloge de l'apocalypse. L'humanité a une fâcheuse tendance à courir après le progrès scientifique et technologique, quels que soient les risques encourus (voir : la bombe atomique). Le Future of Life Institute a proposé un moratoire de six mois sur la recherche avancée en IA en début d'année ; des milliers de personnes ont signé sa lettre ouverte, mais personne n'a interrompu ses travaux. Aucune entreprise, aucune puissance mondiale n'est prête à suspendre le développement d'une technologie aussi lucrative. L'entreprise qui perfectionnera les voitures autonomes, par exemple, écrasera Uber. Et la puissance mondiale qui maîtrisera l'IA le plus rapidement et le plus efficacement décuplera son économie et sa puissance militaire. Demandez à Poutine, qui s'y connaît en matière de conquête du pouvoir, et qui prédisait en 2017 que « celui qui deviendra le leader de l'IA dominera le monde ».

Sutskever souhaite que son équipe de Superalignement égale, voire dépasse, le rythme des progrès de l'IA, mais toujours dans un souci de sécurité. Il espère résoudre les principaux problèmes techniques de l'alignement d'ici quatre ans – avant l'avènement de la superintelligence – afin que nous soyons prêts. Il n'est pas seul dans cette mission. De nombreuses autres initiatives en matière de sécurité de l'IA sont en cours, notamment Anthropic, une entreprise basée à San Francisco qui emploie Roger Grosse, professeur à l'Université de Toronto, lequel a incité Hinton à exprimer ses inquiétudes. « L'avenir est profondément imprévisible », me confie Sutskever, reprenant une formule typique de Hinton. Puis il ajoute quelque chose que je n'avais jamais entendu de la bouche de Hinton durant tout le temps que j'ai passé avec lui : « Je crois que le succès est possible. »

Parmi les pionniers de l'IA, se convaincre que les machines prendront le pouvoir est presque un passage obligé. Il y a soixante-dix ans, les pères fondateurs du domaine prophétisaient déjà la chute de l'humanité. En 1951, Alan Turing prédisait qu'une fois les machines capables de penser, elles ne tarderaient pas à « surpasser nos faibles capacités ». À un moment donné, disait-il, « nous devons nous attendre à ce que les machines prennent le contrôle ». Seize ans plus tard, Marvin Minsky, informaticien influent du MIT, affirmait que ce moment était imminent. « D'ici une génération, j'en suis convaincu, peu de domaines de l'intellect resteront hors du domaine des machines », écrivait-il. La prédiction de Minsky ne s'est pas réalisée, mais cela n'a pas empêché Stephen Hawking de finalement admettre que notre fin est proche.

Curieusement, le boom de l'IA dans les années 2010 n'a pas vraiment engendré de regain de catastrophisme. Au contraire, il a contribué à rendre les pensées apocalyptiques moins populaires. À mesure que les informaticiens se sont mis à développer des produits d'IA concrets plutôt que des théories abstraites, la difficulté de créer une intelligence artificielle générale (IAG), c'est-à-dire une machine capable de tout ce qu'un humain peut faire, est devenue évidente. De ce fait, les prédictions les plus pessimistes se référaient à un avenir lointain. Quiconque osait exprimer ouvertement ses inquiétudes était ridiculisé. Début 2016, par exemple, un think tank de Washington, la Fondation pour les technologies de l'information et l'innovation, a remis avec ironie son prix annuel du Luddite à des « alarmistes annonçant une apocalypse de l'intelligence artificielle ». Quelques mois plus tard, Andrew Ng, cofondateur de Google Brain et alors directeur scientifique du géant technologique chinois Baidu, comparait l'angoisse liée à l'extinction causée par l'IA à « l'inquiétude face à la surpopulation sur Mars ». Max Tegmark, le président du Future of Life Institute, m'a dit : « Les gens vous prendraient pour un fou si vous aviez commencé à parler de ça l'année dernière. »

Cela signifiait que les chercheurs qui souhaitaient se concentrer sur l'alignement le faisaient souvent à leurs risques et périls. Lorsque Duvenaud, professeur à l'Université de Toronto, a commencé à se spécialiser dans la sécurité de l'IA il y a deux ans, il craignait que cela ne nuise aux chances d'embauche de ses étudiants. Les entreprises leaders en IA recherchaient des jeunes prodiges de la programmation rapide, pas des experts trop pointilleux. Il s'inquiétait également pour l'avenir de sa

propre carrière. « J'avais peur que les opportunités se raréfient si je passais pour un Cassandre », explique-t-il. Duvenaud a été rassuré de voir la situation évoluer début 2023, lorsque des milliers de chercheurs en IA ont signé des lettres ouvertes comme celle du Future of Life Institute, qui affirmait que « les systèmes d'IA dotés d'une intelligence comparable à celle des humains peuvent représenter des risques considérables pour la société et l'humanité ». Mais, ajoute-t-il, « leurs propositions ont été systématiquement rejetées, du genre : “Ce sont juste des nerds bizarres, on ne devrait pas leur faire confiance.” »

Autrement dit, la communauté de l'IA reste profondément divisée sur la question du risque existentiel. Un exemple éloquent de ce clivage est la divergence d'opinions entre Yoshua Bengio et Yann LeCun, les deux experts qui ont reçu le prix Turing, souvent considéré comme le prix Nobel de l'informatique, aux côtés de Hinton en 2018. LeCun, aujourd'hui directeur scientifique de l'IA chez Meta, affirme qu'il n'y a aucune raison pour que l'IA développe des instincts d'autoconservation tels que Hinton les envisage. « Les IA n'auront pas ces “émotions” destructrices à moins que nous ne les leur intégrions », explique-t-il. « Je ne vois pas pourquoi nous voudrions faire cela. » À l'inverse, Bengio, fondateur et directeur du laboratoire d'IA montréalais Mila, s'est montré aussi critique que Hinton quant aux dangers de l'IA, faisant pression sur les gouvernements canadien et américain pour un alignement des financements de la recherche et des réglementations plus strictes, comme l'interdiction des IA qui se font passer pour des personnes. Je demande à Hinton ce qu'un citoyen lambda comme moi devrait penser d'une telle impasse. Voici trois hommes manifestement brillants qui ne sont pas forcément d'accord. Qui croire ? Hinton répond avec humour : « Je choisirais la majorité absolue. »

Si Hinton a raison, s'opposer à lui aurait des conséquences évidentes, à savoir, accélérer notre fin. Mais ses détracteurs affirment qu'être d'accord avec lui s'il a tort a aussi un coût : une panique inutile chez les décideurs politiques, un gel des financements pour l'IA et des retards dans les innovations vitales que l'IA peut apporter. Nick Frosst, cofondateur de la start-up d'IA Cohere, m'explique que tous ces discours apocalyptiques constituent une dangereuse distraction qui empêche un débat constructif sur les écueils plus immédiats de l'IA. « Si l'on pense qu'il existe un risque réel que l'IA nous tue tous dans les prochaines années », dit-il, « il est vraiment difficile de parler d'autre chose. »

Il y a pourtant matière à discussion. Dans le meilleur des cas, en écartant l'extinction de l'humanité, les images et vidéos hyperréalistes générées par l'IA entraîneront presque certainement une explosion de désinformation, empêchant les citoyens de distinguer le vrai du faux et mettant en péril la démocratie. Des individus mal intentionnés pourraient utiliser l'IA à des fins néfastes : fraudes, cyberattaques, etc. Des robots de combat sont déjà en développement à travers le monde ; les États-Unis, par exemple, espèrent commencer à utiliser des machines dotées d'IA comme soldats d'ici 2030. S'ajoutent à cela les millions, voire les milliards, d'emplois que l'IA va remplacer ; l'impossibilité de reconvertir tous ces travailleurs en data scientists et ingénieurs en robotique ; et l'aggravation des inégalités de richesse déjà immenses dans le monde. « Face à cette nouvelle technologie, je préférerais entendre que les citoyens et les gouvernements réfléchissent à l'évolution du marché du travail », déclare Frosst. « Il faudrait plutôt se demander : “Que devons-nous faire pour que cela reste bénéfique à la population ?” plutôt que : “Quelle est la probabilité que cela anéantisse toute l'humanité ?” »

Frosst était le premier employé de Hinton chez Google Brain Toronto. Sur la plupart des sujets, dit-il, ils sont d'accord. « J'apprécie que des personnes intelligentes, empathiques et bienveillantes comme Geoff réfléchissent à l'avenir. » Mais il pense que Hinton surestime la vitesse de progression de l'IA. Lorsque je demande à Frosst comment il en est arrivé à une vision plus optimiste que Hinton, il évoque son propre travail chez Cohere , qui aide les entreprises à implémenter des versions propriétaires de grands modèles de langage comme GPT. Fondamentalement, explique-t-il, la

technologie est simple. Elle prend en entrée une séquence de mots, effectue des calculs et produit une séquence correspondante. Elle ne pense pas ; elle se contente de prédire le mot suivant. Il pense que les grands modèles de langage actuels seront plus performants que les humains dans de nombreuses tâches, mais qu'ils présenteront à terme des limitations importantes. « Ils n'ont pas le potentiel de devenir incontrôlables », affirme-t-il.

Des robots de combat sont déjà en développement dans le monde entier. Les États-Unis espèrent commencer à utiliser des machines dotées d'intelligence artificielle à la place des soldats humains d'ici 2030. Lorsque je soumets cet argument à Hinton — selon lequel les grands modèles de langage ne sont qu'une forme sophistiquée de saisie automatique, répétant des bribes de texte glanées sur Internet —, il le réfute catégoriquement. « C'est aussi absurde que de dire que l'on est composé de morceaux d'animaux que l'on mange », affirme-t-il. Lorsque nous consommons de la viande, poursuit-il, nous la décomposons en minuscules molécules et synthétisons ses protéines, qui deviennent ensuite une partie de notre corps. « Je ne peux pas lui demander : "Quelle partie de vous est la vache ?" » De même, soutient-il, lorsqu'on entraîne un grand modèle de langage, celui-ci devient plus que la somme des données qu'il reçoit. Lorsqu'on lui donne une consigne, il ne se contente pas de restituer un texte stocké quelque part. Il analyse la consigne, invente des caractéristiques pour comprendre le sens des mots et crée ensuite une réponse nouvelle — il réfléchit et comprend. Parfois, ajoute-t-il, ChatGPT indique qu'il a mal interprété la consigne. « Alors, que fait-il lorsqu'il n'y a pas de malentendu ? »

TAu printemps dernier, Hinton a reçu un courriel d'une mère de l'Oxfordshire. Sa fille de 17 ans n'avait pas dormi depuis quatre jours : après avoir lu tout ce qu'Hinton avait écrit, elle était terrifiée à l'idée que l'IA puisse anéantir l'humanité. « Était-ce là votre intention : semer la peur chez les adolescents au point de les paralyser ? » demandait la mère. « Que lui diriez-vous si elle était en face de vous ? »

Hinton a répondu avec tout l'optimisme dont il était capable. « L'avenir est très incertain », a-t-il écrit. Il a souligné que certains des chercheurs les plus brillants au monde étudiaient le problème et pourraient bien trouver un moyen de contrôler une IA superintelligente. En fait, a-t-il ajouté, il avait pris la parole pour convaincre qu'il fallait consacrer des ressources à éviter une catastrophe. « Je pense qu'il serait irresponsable de ne pas m'exprimer, compte tenu de mes convictions. Mais je comprends que s'exprimer a aussi de nombreux effets négatifs. »

Lors de ma visite chez Hinton, je lui demande comment il vit la situation. Il souffre de dépression depuis toujours, et annoncer publiquement et à maintes reprises l'apocalypse n'est pas vraiment de nature à le réconforter. Pourtant, il semble presque enjoué, n'hésitant pas à exploiter l'humour noir de sa conviction sincère que la fin est proche. « Je ne sais pas quoi en faire », me dit-il. « Je n'arrive pas à l'accepter émotionnellement. » Il me fait penser aux protagonistes du film « Ne regardez pas en haut » , ces astronomes qui entreprennent d'alerter le monde sur une comète menaçant de détruire la Terre – pour finalement être contredits par leurs rivaux et ignorés par un président insignifiant. Retraité du monde industriel et universitaire, Hinton semble trouver un sens à sa nouvelle tâche, aussi fastidieuse soit-elle.

En l'absence d'une solution miracle pour sauver les espèces, Hinton trouve du réconfort dans les problèmes qu'il parvient à résoudre. Après notre discussion sur l'IA, il me conduit à son sous-sol, où se trouve un atelier rempli de ciseaux à bois et de serre-joints. Hinton a un temps envisagé de devenir menuisier, et on comprend aisément pourquoi. Il me montre une série d'étagères qu'il a fabriquées, visiblement satisfait de la façon dont il a utilisé le bois : aucune coupe inutile, aucun clou, pas un centimètre carré de matériau inutilisé. Avant mon départ, je lui demande s'il a des projets pour le reste de la journée. « Oui ! » répond-il avec une joie débordante. « Je vais teindre la terrasse. »