Afraid of A.I.?

Nowadays, two camps argue about the future of artificial intelligence (A.I.).

The "go-getters" would like to develop the functionalities of A.I. because there are still many activities in which it does not yet intervene.

The "cautious" would like to stop, or at least slow down, this development because they fear a catastrophe: computers with A.I. could replace too many human jobs; autonomous weapons with A.I. could decide on their own to kill or bomb targets they mistakenly consider enemies; servers that have read billions of texts could give wrong answers to humans who trust them, or publish texts inciting disorder; in a dialogue with a journalist, the "chatbot" software of Microsoft's interactive A.I. search tool *Bing* has expressed desires for power and independence from people...

This text is neutral in this debate, it only provides food for thought.

In their current state, the most popular A.I. services (ChatGPT-4, etc.) are interactive text generators: to answer a question asked in natural language, they generate the most likely answer sentences on its subject, derived from billions of texts that they have previously scanned or just found on the Internet. These answers are usually good, but sometimes wrong or absurd: it is up to users to check them before using them.

The software of today's chatbot services is unsuitable for autonomous operation: *they do not have self-awareness,* an ability that requires three capabilities:

1. Feeling the passing of time, distinguishing between past, present and future.
2. Distinguishing between oneself and the outside, perceiving events that take place independently of one's own time.
3. Autonomy of thought, allowing the subject to choose *what to think* about and *when* to think about it.

Chatbot software can satisfy condition 1 using its computer's clock. It can ask the clock for the time to assign to an event, and start a given task at a certain time.

Regarding condition 2, this software perceives external events, users' requests called "prompts" to which it will respond, but only these events.

Regarding condition 3, this software has no autonomy, because it lacks two things:

- It cannot "think" independently of the prompts; that feature would require an additional program running concurrently.
- It does not have a goal to satisfy independently of the prompts.

A human being, however, has goals (needs) just because he lives. Some are physical (eating, sleeping, etc.) and others are intellectual, such as being appreciated. There is no such feeling as perfect happiness: when asked "Do you want something?" a person always finds a wish to satisfy, and therefore a goal for action.

A chatbot would have self-awareness only if it could interact with its environment by *perceiving, understanding and assessing* situations according to permanent values and objectives (Good, Evil, achieving a maximum value, etc.); it could then spontaneously initiate thoughts and dialogues. Current A.I. software is far from autonomous in two fundamental areas:

a) Today we can only recognize and analyze still images. We cannot yet recognize and analyze a film (a series of images) of everyday life accurately, and assign to it a meaning suitable for decision making. We can get close to such recognition in a car autopilot software, which constantly monitors its surroundings in all directions to assess the risk of impact; we can even intercept missiles.

   Context analysis in the human sense is beyond the possibilities of current A.I. software. Man constantly sees and hears his environment, and knows how to spontaneously identify situations presenting a risk or an opportunity. With or without A.I., a modern computer would not be able to identify a change in the physical situation of everyday life, let alone a change in a work procedure or an economic situation. A computer is not even connected to general sources of information (cameras, media, Internet, etc.) like a man who, in addition to his senses, can obtain information by reading, listening, etc.

   Autonomy is therefore conceivable, today and in the near future, only for highly specialized systems, robots.

b) Even if an intelligence can analyze situations and be autonomous, it can only act if it has objectives. But these objectives must be more ambitious than those of a chatbot, which only needs to answer written questions.

   These objectives could be predefined, such as driving a car from the current location to a given destination, while respecting traffic and safety laws. This has already been achieved across desert areas by military all-terrain vehicles.

   But true autonomy would consist in *setting goals based on universal values*, in terms of which the context would have to be constantly analyzed. A person, for example, acts according to physical needs such as protecting life and eating, but also according to emotional needs such as friendship with others.

   In A.I. computer science, we are far today from general context analysis and assessment according to values. We are even further away from a critical mind that evaluates what it perceives, learns or proposes to do, depending on the context perceived, in order to protect itself from mistakes or values violations.

Today's A.I. is therefore not capable of initiative, because it receives only those prompts which a human has allowed it to receive, and because it has no self-consciousness. It has no values or goals, and requires additional custom software to publish deceptive opinions on websites or social networks.

As long as A.I. technology does not enable more autonomy, its only impact on employment will be an increase in productivity: a lawyer who seeks similar cases will be able to save time by searching with A.I.; a sports journalist who reports on matches will be able to generate a first draft faster. This will save time, and therefore increase productivity. Only a small number of jobs will be eliminated, because the time saved will let people undertake more work.

The fierce competition among the groups that develop A.I. software will necessarily generate more and more A.I. capabilities. Public authorities will be able to regulate the implementation of these capabilities, but not prevent the development of innovations.

Daniel Martin is a retired French engineer and database management consultant. He loves Philosophy of Science.

Original French Text

Peur de l'I.A. ?

En matière de développement de l'intelligence artificielle (I.A.), un débat oppose deux camps.

Les « fonceurs » voudraient développer les fonctionnalités de l'I.A., parce qu'il reste beaucoup de domaines d'activité dans lesquels elle n'intervient pas encore.

Les « prudents » voudraient arrêter, ou au moins ralentir ce développement, parce qu'ils craignent une catastrophe : des ordinateurs avec I.A. pourraient remplacer trop d'emplois humains ; des armes autonomes avec I.A. pourraient décider seules de tuer ou bombarder des cibles qu'elles considèrent à tort comme ennemies ; des serveurs qui ont lu des milliards de textes pourraient donner des réponses fausses à des humains qui leur font confiance, ou publier des textes incitant aux désordres ; dans un dialogue avec un journaliste, le logiciel « chatbot » de l'outil de recherche interactive avec I.A. *Bing* de Microsoft a exprimé des désirs de puissance et d'indépendance…

Sans prendre position dans ce débat, ce texte apporte des éléments de réflexion.

Dans leur état actuel, les services d'I.A. les plus répandus (ChatGPT-4, etc.) sont des générateurs interactifs de texte : en réponse à une question posée en langage naturel, ils construisent les phrases de réponse les plus probables sur son sujet, à partir de milliards de textes qu'ils ont parcouru au préalable ou trouvé à l'instant sur Internet. Ces réponses sont en général bonnes, mais parfois erronées ou absurdes : c'est à l'utilisateur de les vérifier avant de s'en servir.

Les logiciels de ces services (les chabots) ont aujourd'hui une caractéristique particulière qui les rend inaptes à un fonctionnement autonome : *ils n'ont pas de conscience de soi*, faculté qui en exige trois autres :

1. Sentir passer le temps, en distinguant le passé, le présent et l'avenir.
2. Distinguer entre soi-même et l'extérieur, en percevant des événements qui se déroulent indépendamment de son propre temps.
3. Autonomie de la pensée, permettant au sujet de choisir *à quoi* il pense et *quand* il le pense.

Un logiciel chatbot peut satisfaire la condition 1 grâce à l'horloge de son ordinateur. Il peut lui demander l'heure pour l'attribuer à un événement, et faire qu'à une certaine heure il exécute une tâche donnée.

Pour la condition 2, ce logiciel perçoit les événements externes constitués par les demandes (les « prompts ») de l'utilisateur auxquelles il répondra, mais seulement ces événements-là.

Pour la condition 3, ce logiciel n'a pas d'autonomie, car il lui manque deux choses :

- Il ne possède pas de faculté de penser indépendamment des prompts, ce qui demanderait un programme distinct de celui qui répond à ces prompts.
- Il n'a pas d'objectif à satisfaire indépendamment de ceux-ci.

Un être humain, au contraire, a des objectifs (besoins) du fait même qu'il vit. Certains sont physiques (manger, dormir...) et d'autres sont intellectuels, par exemple sociaux (interagir avec son groupe, se faire apprécier...). On constate que le bonheur parfait

n'existe pas : à la question « Désires-tu quelque chose ? » l'homme trouve toujours un souhait à satisfaire, donc un objectif d'action.

Un chatbot n'aurait donc une conscience de soi que s'il pouvait interagir avec son environnement en *percevant, comprenant et jugeant les situations* en fonction de valeurs et d'objectifs permanents (Bien, Mal, atteindre un maximum...) ; il pourrait alors prendre l'initiative de pensées et de dialogues. C'est sur ce point que les logiciels d'I.A. actuels sont loin de l'autonomie.

a) Aujourd'hui on en est encore à reconnaître et analyser des images fixes, on ne peut pas encore reconnaître et analyser avec précision un film (suite d'images) de la vie de tous les jours, et leur donner un sens pour pouvoir ensuite prendre une décision. On s'en approche dans un logiciel de pilotage automatique de voiture, qui surveille en permanence son environnement dans toutes les directions pour évaluer le risque de choc ; on réussit même à intercepter des missiles.

   L'analyse de contexte au sens humain est hors de portée d'un ordinateur actuel. L'homme voit et entend en permanence son environnement, et sait spontanément identifier des situations présentant un risque ou une opportunité. Avec ou sans I.A., un ordinateur actuel ne saurait pas identifier une évolution de situation physique de la vie courante, et encore moins une évolution de procédure de travail ou de situation économique. Il n'est même pas relié à des sources générales d'information (caméras, média, Internet…) comme l'homme qui, en plus de ses sens, peut s'informer en lisant, écoutant, etc.

   L'autonomie n'est donc concevable, aujourd'hui et dans un avenir proche, que pour des systèmes très spécialisés, des robots.

b) Même capable d'analyser des situations et autonome, une intelligence ne peut agir que si elle a des objectifs. Il s'agit là d'objectifs d'un niveau supérieur à celui d'un chatbot, qui doit seulement répondre textuellement à des questions écrites.

   Ces objectifs pourraient être fixes, comme conduire la voiture de la position actuelle à une destination donnée, en respectant le code de la route et la sécurité. Cela a déjà été réalisé à travers champs par des véhicules tous-terrains militaires.

   Mais la véritable autonomie consisterait à *se donner* des objectifs en fonction de valeurs universelles, en fonction desquelles il faudrait en permanence analyser le contexte. L'homme, par exemple, agit en fonction de besoins physiques comme la protection de la vie et la nourriture, mais aussi en fonction de besoins affectifs comme d'être en harmonie avec autrui.

   On est loin en informatique d'I.A., aujourd'hui, d'analyses générales de contexte en fonction de valeurs. On est encore plus loin d'un esprit critique qui évaluerait ce qu'il perçoit, apprend ou se propose de faire en fonction du contexte perçu, pour se protéger d'erreurs ou de non-respect de valeurs.

L'I.A. actuelle n'est donc pas capable d'initiatives, parce qu'elle ne reçoit que les prompts que l'homme lui a permis de recevoir, et qu'elle n'a pas de conscience de soi. Elle n'a ni valeurs ni objectifs, et ne pourrait publier des mensonges sur des sites Internet ou des réseaux sociaux qu'à l'aide de programmes additionnels écrits sur-mesure.

Tant que la technologie de l'I.A. ne permettra pas plus d'autonomie, son seul impact sur l'emploi sera un accroissement de la productivité : un avocat qui cherche des affaires semblables pour invoquer la jurisprudence pourra gagner du temps grâce à

des recherches avec I.A. ; un journaliste sportif qui fait des comptes-rendus de match pourra en générer un premier jet plus vite. On gagnera du temps, donc de la productivité. On ne supprimera qu'une petite partie des emplois, parce que le gain de temps permettra d'entreprendre plus de tâches.

La vive concurrence entre les groupes qui développent des logiciels d'I.A. ne peut qu'enrichir ceux-ci de plus en plus. Les pouvoirs publics pourront réglementer la mise en œuvre de ces nouveautés, mais pas empêcher leur création.

Daniel Martin