Chers membres du réseau de l'ancienne FSFA, chers intéressés,

Est-il possible que des systèmes d'IA autonomes, produisent des effets indésirables – non programmés et autodidactes – et ne respectent pas les directives, les lignes rouges (problème d'alignement) ? Il existe des informations inquiétantes à ce sujet. Mais ce qui est particulièrement alarmant, c'est qu'il existe certains dilemmes qui limitent les principaux acteurs dans leur champ d'action. Il est grand temps d'en débattre – et pas seulement dans le domaine de l'éducation.

Des systèmes d'IA dangereux – bien qu'entraînés avec des directives éthiques ?
Des chercheurs du MIT (Massachusetts Institute of Technology) ont étudié le jeu Cicero de Meta et ont constaté que l'IA, bien qu'entraînée à être honnête et à ne jamais tromper délibérément les humains, s'apprenait elle-même des trucs et des tromperies afin d'atteindre l'objectif fixé dans l'algorithme, à savoir gagner. Le Courthouse News Service, un service d'information américain, en a rapporté les faits début mai 2024, ici.

Hélas, s'ils agissaient de manière autonome...
Les résultats d'une étude scientifique de l'université de Stanford sur les propositions de différents chatbots d'IA dans des situations de guerre simulées ne sont peu rassurants, ici. Tous les modèles – examinés notamment GPT-3.5, GPT-4 d'OpenAI, Claude 2 d'Anthropic et Llama 2 de Meta – ont montré des schémas d'escalade difficilement prévisibles. Certains recommandant une attaque nucléaire, bien que les outils aient été rendus attentifs aux conséquences possibles dans le monde réel. Ce phénomène est connu dans la recherche sur l'alignement sous le nom de « power-seeking » pour atteindre les objectifs. Et si ces outils ne servaient pas à la création de stratégies par nous, les humains, mais agissaient comme des systèmes d'IA autonomes ?

Le dilemme des géants de la technologie – qui se méfient les uns des autres ?
Il est réjouissant de constater que 16 géants de l'IA ont décidé, lors du Sommet de l'IA des 21 et 22 mai 2024 à Séoul, d'élaborer des directives pour les tests de sécurité de l'IA avec des barrières de risque, ici. Le Sommet de l'IA fait suite à celui de Londres, qui a conduit à la déclaration de Bletchley. Nous attendons des effets – avant même le prochain sommet qui se tiendra en France. Se pourrait-il que les géants de la technologie sachent que des directives de sécurité sont nécessaires, mais qu'aucun ne veuille freiner plus que les autres sa propre force d'innovation ? Et qui contrôle le respect des règles ? Le premier qui bouge, qui s'en tient trop strictement aux directives, a déjà perdu dans la compétition pour l'innovation. On ne peut guère compter sur l'autocontrôle des géants de la technologie. Une responsabilité civile ou un système d'amendes pourraient avoir un effet de frein.

Le dilemme des organisations étatiques – la dimension géopolitique
Les organisations étatiques pourraient-elles faire face aux géants internationaux de la technologie et le veulent-elles ? La compétition des géants de la technologie en matière d'IA a également une dimension géopolitique. Depuis longtemps, les Etats-Unis mettent tout en œuvre, avec les sanctions technologiques contre la Chine, pour assurer leur avance, par exemple dans la production de chips, importante pour le développement de l'IA. Pékin a réagi par des investissements records. Une réglementation étatique trop précoce pourrait désavantager le propre développement de l'IA par rapport à la concurrence. Le potentiel d'innovation des systèmes d'IA est énorme et déterminant pour l'avenir des économies nationales. Les États-Unis ne veulent pas le freiner. Quel pourrait être le rôle de l'UE, qui est à la traîne en matière de développement de l'IA, mais qui dispose d'un potentiel de pression en tant que grand marché ? La compétition géopolitique entre les Etats-Unis et la Chine est aussi une concurrence entre démocraties et autocraties. C'est pourquoi l'UE n'ira pas de l'avant.

Une course contre la montre
Il serait judicieux de créer au sein de l'ONU une autorité scientifique et technique dotée d'un statut autonome qui – similaire à l'Agence internationale de l'énergie atomique – serait chargée de surveiller le développement de systèmes d'IA toujours plus puissants et serait autorisée à imposer de manière proactive des règles de sécurité convenues au niveau international lorsque certaines étapes sont atteintes. En effet, le développement d'une IAG, avant même qu'elle ne soit lancée sur le marché, pourrait déjà déclencher des processus dangereux et irréversibles. La compétition entre les géants de la technologie accélère le développement de systèmes d'IA dangereux. Les processus politiques, notamment dans les démocraties et lors de l'élaboration d'accords internationaux, prennent du temps.

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller