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

Des outils d'IA toujours plus puissants sont développés à un rythme accéléré. On assiste à de véritables percées vers l'IAG (Intelligence Artificielle Générale), des systèmes d'IA qui poursuivent de manière autonome les objectifs fixés. Dans ce courriel et dans le suivant, nous mettons en lumière certains aspects inquiétants qui devraient être un sujet de discussion important dans l'éducation.

Les alertes à l'IAG – juste avant la conférence sur la sécurité de l'IA à Séoul
Peu avant la conférence sur la sécurité qui se tenait à Séoul les 21 et 22 mai 2024, 25 coryphées de l'IA, parmi lesquels les lauréats du Turing Award 2018, Yoshua Bengio et Geoffrey Hinton, ont alerté dans la célèbre revue Science (article entier payant) au sujet d'un développement incontrôlé d'une IAG. Selon eux, les entreprises d'IA investissent des sommes énormes dans des outils toujours plus performants et bien trop peu dans la recherche nécessaire sur les questions de sécurité. Il y a un risque que des systèmes d'IA autonomes soient développés qui dépassent les lignes rouges que nous avons fixées. Cela pourrait se produire par inattention en raison d'une programmation erronée ou parce que les systèmes d'IA, en apprenant par eux-mêmes, ignorent les directives humaines, y compris les mécanismes de désactivation. Les conséquences seraient irréversibles. Les auteurs évoquent le risque de nous marginaliser, voire de faire disparaître l'humanité. L'équipe d'auteurs exige des règles de sécurité proactives, convenues au niveau international, qui entreraient automatiquement en vigueur lorsque certaines étapes du développement futur de l'IA seraient atteintes.

Une pause de réflexion ? – De nombreuses voix critiques
L'avertissement selon lequel une IA surhumaine pourrait conduire à l'extinction de l'humanité n'est pas nouveau. En mars 2023 déjà, une lettre ouverte (ici), signée par plus de 26 000 personnes à l'époque (plus de 33 000 aujourd'hui), demandait une pause de réflexion de six mois dans l'entraînement de nouveaux modèles d'IA. En réaction, de nombreuses voix critiques se sont élevées, tant au niveau national (Prof. Dr Andreas Krause et Dr Alexander Ilic du ETH AI Center, ici) qu'international (DAIR Institute, une organisation américaine à but non lucratif, ici). Voici quelques critiques. Un moratoire serait-il même contrôlable et applicable ? Certains veulent-ils continuer à faire de la recherche en secret pour rattraper le retard ? Une histoire apocalyptique est-elle destinée à détourner l'attention des dangers potentiels actuels ? Il serait plus ciblé d'exiger la transparence de la part des géants de la technologie et de les obliger à se responsabiliser.

Directives pour les tests de sécurité de l'IA
Depuis la publication de la lettre ouverte, la compétition en matière d'IA s'est accélérée. Parmi les premiers signataires – outre d'éminentes capacités de recherche en IA – se trouvait également Elon Musk, qui veut entre-temps construire le plus grand supercalculateur d'IA (ici) pour sa start-up xAI, fondée le 9 mars 2023. Au moins, les géants de la technologie semblent être conscients des risques de sécurité liés au développement d'outils d'IA toujours plus performants. Lors du Sommet AI des 21 et 22 mai 2024 à Séoul, 16 entreprises d'IA ont signé un document selon lequel des directives pour les tests de sécurité de l'IA avec des barrières de risque doivent être élaborées, ici.

Une IA agissant de manière autonome et se tournant contre les humains ?
Est-il vraiment concevable qu'une IAG telle que les géants de la technologie veulent la développer dans la compétition ne puisse plus être contrôlée et ne puisse plus être arrêtée dans sa force destructrice ? Jürgen Schmidhuber, directeur de l'IDSIA (institut de recherche sur l'IA au Tessin), qualifié de « père de l'IA avancée » (dans le New York Times, ici), est lui aussi convaincu qu'avec l'IAG, des acteurs très supérieurs à nous, les humains, verront le jour. Il ne voit toutefois pas de danger existentiel pour l'humanité, car il n'y a aucun motif de nous exterminer uniquement parce que nous sommes plus bêtes, ici. Vraiment sans danger ? Il y aura quelques informations inquiétantes à ce sujet dans le prochain e-mail de demain vendredi.

Incontestable – le potentiel de danger dans les mains d'individus destructeurs
Tout le monde est d'accord sur ce point. Le poète italien Italo Svevo a thématisé le danger potentiel des inventions techniques dans son œuvre « La coscienza di Zeno », publiée en 1923. L'un d'eux vole et met le feu à l'explosif d'un inventeur génial et fait sauter le Monde («…e la terra ritornata alla forma di nebulosa errerà nei cieli priva di parassiti e di malattie»).

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

Nouveau droit de la protection des données : Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter !