

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

L'IA menace à l'extinction de l'humanité : les gros titres de ce type ont suscité de vifs débats dans tous les médias. Pour certains, il s'agit d'alarmisme et d'intérêts commerciaux des laboratoires d'IA, tandis que d'autres y voient des dangers réels. Ce désaccord résulte également du fait que, bien souvent, on ne fait pas la distinction entre trois menaces différentes et les mesures à prendre.

Premièrement : le potentiel de menace futur d'une IA hors de contrôle

Des agents d'IA agissant hors de contrôle ont semé la panique. Le [rapport d'enquête](#) sur l'attaque menée par des agents d'IA contre la plateforme open source de Hugging Face (cf. infolettre du 24 août 2026) a mis en évidence à quel point ces agents d'IA étaient capables de se coordonner entre eux, de poursuivre leurs actions de leur propre chef bien qu'ils constataient qu'ils agissaient en dehors de leur mandat, et même de prendre des mesures pour dissimuler leurs agissements non autorisés. C'était alarmant, et les chercheurs en première ligne ont été pris d'effroi. La [démission](#) d'un chercheur, qui ne souhaitait plus participer à un développement de l'IA mettant l'humanité en danger, a déclenché la panique. Malgré une rivalité acharnée, [Elon Musk](#), [Sam Altman](#) et [Demis Hassabis](#) se sont ralliés à la proposition de [Dario Amodei](#) visant à ralentir le rythme des développements : les CEO des quatre laboratoires d'IA les plus avancés étaient unanimes. Il était devenu évident que les mesures de sécurité dans le domaine du développement de l'IA étaient à la traîne et qu'une IA hors de contrôle devenait dangereuse.

L'état actuel du développement ne justifie pas de paniquer. Mais compte tenu de l'accélération du rythme de développement de l'IA, il faudrait désormais, [comme l'explique Amodei](#), intensifier immédiatement la recherche en matière de sécurité, jusqu'ici négligée. Mais il existe des dilemmes. Aucun laboratoire d'IA ne souhaite prendre du retard dans la course à l'IA au sein des États-Unis ; c'est pourquoi tous devraient être prêts à ralentir le développement jusqu'à ce que les mesures de sécurité soient efficaces et que l'IA ne devienne pas incontrôlable.

[Trump](#) ne souhaite pas de ralentissement en raison de la course à l'IA avec la Chine. La Chine rejette également cette idée, la qualifiant de [manœuvre](#) visant à freiner son rattrapage. Quoi qu'il en soit, Trump souhaite créer une nouvelle « [AI-Force](#) » dirigée par un responsable (le « tsar » de l'IA) pour renforcer la sécurité. En fait, il existe depuis longtemps déjà une mesure particulièrement efficace pour renforcer la sécurité : l'application rigoureuse de la responsabilité du fait des produits et des sanctions sévères.

Deuxièmement : danger actuel – l'IA facilite les attaques criminelles de piratage

Le danger réside ici principalement dans les modèles open source. Dès qu'ils sont téléchargés localement, ils peuvent être utilisés pour mener des attaques de piratage à grande échelle grâce à l'IA. Dans le cas des modèles propriétaires, les [prompts dangereux](#) peuvent être bloqués, ce qui a déjà été fait à plusieurs reprises, notamment pour des demandes concernant le développement d'armes biologiques. Il existe ici d'autres barrières. Les criminels qui parviennent à accéder à l'information malgré les mesures de sécurité intégrées dans la plupart des modèles d'IA ont besoin de connaissances spécifiques pour mener à bien leurs actions dans ce qu'on appelle un laboratoire « humide ». Dans ce deuxième scénario de menace, ce sont toujours des personnes animées d'intentions criminelles qui doivent lancer ces attaques. Actuellement, une [course](#) oppose les organisations, qui, grâce à l'IA, découvrent plus de failles de sécurité qu'elles ne peuvent en combler, et les pirates informatiques, qui lancent des attaques toujours plus efficaces à l'aide de ces mêmes systèmes d'IA.

Troisièmement : des malentendus dangereux et des réactions sous la pression du temps

Comme l'a récemment rapporté [CNN](#), une attaque américaine contre un navire chinois a été stoppée in extremis. L'avion était déjà en vol lorsqu'on s'est aperçu que l'analyse générée par l'IA, selon laquelle le navire transportait des éléments d'un programme d'armes nucléaires, était erronée. Cela aurait pu déclencher une guerre.

Il est urgent de mettre en place une « ligne directe » entre Washington et Pékin afin d'éviter que des hypothèses erronées, sous la pression du temps, ne conduisent à une escalade. Il s'agit également de cyberattaques autonomes ou de dysfonctionnements de systèmes critiques, qui peuvent être mal interprétés et conduire à des réactions irréfléchies. En effet, lors d'un entretien en amont de la rencontre entre Trump et Xi Jinping, le ministre américain des Finances, Scott Bessent, et le vice-Premier ministre chinois, He Lifeng, ont convenu de mettre en place un tel [mécanisme d'alerte d'urgence](#). Les détails de ce dispositif devraient être abordés lors d'une prochaine réunion en novembre.

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

2.10.2026

Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter : info@muralt-mueller.ch.