INSTITUT Montaigne

institutmontaigne.org
NAC2304-01

NOTE D'ACTION - Avril 2023

Investir l'IA sûre et digne de confiance : un impératif européen, une opportunité française

Alors que les systèmes d'intelligence artificielle (IA) ont jusqu'à présent été très spécialisés, capables uniquement d'effectuer un nombre de tâches restreint, nous assistons désormais à un changement de paradigme. **Le développement rapide de l'IA - incarné par les systèmes d'IA à usage général comme ChatGPT, capables d'effectuer un grand nombre de tâches - présage de bouleversements technologiques considérables.** L'IA pourrait ainsi rapidement constituer un facteur de compétitivité décisif pour les entreprises comme pour les pays. En témoigne la croissance exponentielle des montants qui y sont investis : 92 milliards de dollars d'investissement privé en 2022, soit près de 20 fois plus qu'il y a dix ans.

Un enjeu de sécurité et un impératif de sûreté

L'accélération du développement de l'IA constitue néanmoins un enjeu de sécurité majeur et croissant. À l'image de la technologie nucléaire, l'IA est une technologie intrinsèquement duale qui peut rapidement être détournée à des fins malveillantes. Les systèmes d'IA d'apprentissage automatique, qui apprennent à effectuer une tâche à partir d'exemples plutôt que de règles prédéfinies, posent également un risque de sûreté et de défaillance inédit, puisque leur nature statistique les rend intrinsèquement imprévisibles. Ils sont ainsi peu robustes, c'est-à-dire que leur comportement peut subitement changer dans des environnements nouveaux, et sont difficilement explicables : ce sont des "boîtes noires" qui fonctionnent en autonomie, sans que l'on sache réellement ni comment ni pourquoi. Ainsi, Google (BARD),Microsoft (Bing) ou OpenAI (ChatGPT) ne parviennent à prévenir ni les erreurs factuelles ni les dérives violentes ou biaisées de leurs agents conversationnels. Au fur et à mesure qu'ils gagnent en capacité et en autonomie, la bonne spécification des objectifs de ces systèmes, c'est-à-dire leur alignement avec l'intérêt général, devient dès lors un enjeu majeur. **Avec les avancées rapides de l'IA et sa dissémination massive dans l'ensemble des secteurs d'activité, ce risque de défaillance pourrait rapidement augmenter et représenter un enjeu de sûreté et de résilience aux échelles nationale et internationale.**

Alors que les systèmes d'intelligence artificielle (IA) ont jusqu'à présent été très spécialisés, capables uniquement d'effectuer un nombre de tâches restreint, nous assistons désormais à un changement de paradigme. **Le développement rapide de l'IA - incarné par les systèmes d'IA à usage général comme ChatGPT, capables d'effectuer un grand nombre de tâches - présage de bouleversements technologiques considérables.** L'IA pourrait ainsi rapidement constituer un facteur de compétitivité décisif pour les entreprises comme pour les pays. En témoigne la croissance exponentielle des montants qui y sont investis : 92 milliards de dollars d'investissement privé en 2022, soit près de 20 fois plus qu'il y a dix ans.

Définir nos préférences pour les systèmes d'IA : un enjeu sociétal

Le développement de systèmes d'IA avancés et leur déploiement à grande échelle constituent également un enjeu sociétal de premier ordre. En effet, dans la mesure où les systèmes d'IA prennent des décisions ou effectuent de recommandations, elles sont porteuses de valeurs et impactent nécessairement notre liberté, que leurs concepteurs le veuillent ou non. **Les laboratoires d'IA qui développent des modèles d'IA à usage général tentent désormais d'améliorer la performance de ces systèmes en y intégrant explicitement les préférences humaines, c'est-à-dire un modèle de valeurs.** ChatGPT s'appuie par exemple sur l'apprentissage par renforcement à partir de retours humains (RLHF). Interpréter les préférences humaines en nous demandant notre avis (RLHF par exemple), en observant nos comportements ou en spécifiant une liste de principes moraux implique un parti pris philosophique et une réflexion éthique approfondie avant tout déploiement opérationnel au cœur de nos machines. **Or la spécification des préférences humaines pour des systèmes d'IA constitue un domaine de recherche naissant, stratégique mais quasiment inexploré à date.**

L'IA sûre et digne de confiance : un avantage stratégique sur lequel miser

Grâce à des moyens accrus - notamment privés - les États-Unis et la Chine ont acquis une longueur d'avance substantielle en matière de développement économique et technologique de l'IA. L'Europe a ainsi cumulé un retard difficilement rattrapable. **Miser sur l'IA sûre et digne de confiance constitue dès lors notre meilleure stratégie de différenciation pour se positionner en acteur clé de l'IA. Elle constitue par ailleurs un impératif pour protéger notre sécurité et notre modèle de société.**

La sûreté et la confiance constituent désormais une barrière technologique importante au développement des systèmes d'IA à usage général et une préoccupation centrale des meilleurs talents internationaux de l'IA. L'Union européenne s'apprête par ailleurs à imposer aux systèmes d'IA des exigences de sûreté et de confiance grâce à une réglementation extraterritoriale, couplée à une directive en matière de responsabilité civile pour l'IA et à un travail pionnier sur les normes. Ce cadre réglementaire pourrait avoir la même portée internationale que le RGPD avant lui.

Au sein de l'Europe, la France est particulièrement motrice et s'est positionnée comme *leader* sur le sujet. Elle est notamment à l'origine de l'inclusion des modèles d'IA "à usage général" (type ChatGPT) dans le règlement européen sur l'IA et se démarque comme étant motrice dans les efforts de normalisation menés au niveau européen. Surtout, elle recense une expertise mondiale sur plu-

sieurs briques techniques clés pour développer des systèmes d'IA à usage général sûrs et dignes de confiance : en recherche fondamentale, grâce à des chercheurs de rang mondial en mathématiques et en IA capables d'attirer les meilleurs talents internationaux ; en ingénierie système et logicielle pour la sûreté, grâce à un Grand Défi sur l'IA de confiance pour les systèmes critiques et à un écosystème d'industriels ; en développement de grands modèles d'IA à usage général, grâce au projet Bloom, un grand modèle de langage comme ChatGPT développé avec des chercheurs français et les puissants ordinateurs du Centre national de la recherche scientifique (CNRS).

Si la France et l'Europe souhaitent pleinement capitaliser sur cette opportunité inédite, elles doivent adopter une approche ambitieuse pour développer des systèmes d'IA à usage général véritablement sûrs et dignes de confiance d'une part, et pour réguler les systèmes d'IA à usage général dangereux d'autre part.

Objectif 1 : Faire de la France un leader mondial de la R&D dans la sûreté et la confiance des modèles d'IA à usage général

Recommandation 1 :
Attirer en France les meilleurs chercheurs internationaux de l'IA avec un appel porté au plus haut niveau de l'État, sur le modèle de l'initiative "Make Our Planet Great Again", centré sur le développement de systèmes d'IA à usage général sûrs et dignes de confiance.

Recommandation 2 :
Mener un projet d'innovation de rupture pour développer des systèmes d'IA à usage général sûrs et dignes de confiance, doté de 100 millions d'euros et d'une gouvernance agile, qui s'appuie sur les forces de l'écosystème français.

Recommandation 3 :
Créer un pôle de recherche mondial sur la compréhension des préférences humaines et leur bonne spécification pour des systèmes d'IA à usage général. Confier la coordination de ce pôle à un institut de recherche emblématique (ENS ou 3IA par exemple) et assurer son financement via une enveloppe dédiée, par exemple des Programmes et équipements prioritaires de recherche (PEPR).

Recommandation 4 :
Faire de l'IA sûre et digne de confiance un projet important d'intérêt européen commun (PIIEC) permettant d'assouplir les règles d'aides d'État et/ou l'un des "produits phares" de l'Union européenne dotés d'environ 1 milliard d'euros.

Recommandation 5 :
Développer en France deux référentiels (*benchmarks*) pour la recherche permettant de mesurer la confiance et la performance d'un système d'IA à usage général.

Recommandation 6 :
Créer une discipline de sûreté de l'IA (ou génie de l'IA) en conditionnant le financement public des formations à l'IA à l'intégration d'un module sur la sûreté et la confiance de l'IA.

Objectif 2 : Définir un cadre réglementaire européen pour la sûreté et la confiance de l'IA à usage général et favoriser son adoption dans le monde

Recommandation 7 :

Concrétiser la proposition de la France d'inscrire les systèmes d'IA à usage général dans la réglementation européenne de l'IA et favoriser son adoption dans le monde via le E.U.-U.S. *Trade and Technology Council* (TTC) et le G20.

Recommandation 8 :

Confier au futur régulateur français de l'IA une expérimentation pilote ou un audit à blanc du processus d'audit de l'IA prévu par la réglementation européenne, afin d'accompagner la montée en puissance d'un écosystème d'audit français (entreprises, auditeurs, régulateur).

Recommandation 9 :

Développer au sein du futur régulateur français de l'IA et en association étroite avec les acteurs de l'évaluation comme le LNE un "bac à sable" (*sandbox*) réglementaire de l'IA, pour tester sans conséquence juridique le degré de conformité de nouveaux systèmes d'IA et d'IA à usage général.

Recommandation 10 :

Confier au futur régulateur français de l'IA la création d'une base de données de référence de documentation des défaillances de systèmes d'IA.