

Fresque de l'IA



0

Données d'entraînement



2

LOT 1

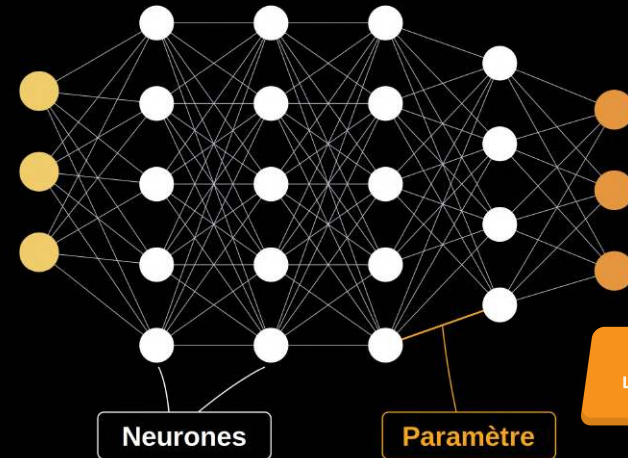


Modèle d'Intelligence Artificielle (IA)

Entrée
(ex: texte)

Couches cachées

Sortie
(ex: texte)



1

LOT 1



Amélioration des capacités



3

LOT 1



Modèle d'Intelligence Artificielle (IA)

Un modèle d'IA est un programme informatique qui apprend à partir d'une grande quantité de données à atteindre un objectif fixé. Une fois entraîné, il peut produire des réponses, des contenus (images, vidéos, sons) ou prendre des décisions pour accomplir son objectif.

LOT 1 

Fresque de l'IA

Adaptée de la Fresque de la sécurité de l'IA du Centre pour la Sécurité de l'IA (CeSIA), elle-même inspirée de la Fresque du Climat, la Fresque de l'IA par Pause IA est un atelier collaboratif pour comprendre ensemble les enjeux de l'IA, sans prérequis technique.

Les participant·es relient les cartes pour dresser une vue d'ensemble : comment l'IA fonctionne, ce qu'elle sait faire, ses impacts, ses risques et les solutions disponibles.

Amélioration des capacités

Pour entraîner un modèle d'IA plus performant, il faut mobiliser des moyens considérables : davantage de puissance de calcul, des équipes de recherche spécialisées et d'immenses volumes de données d'entraînement. Plus l'investissement est important, plus il devient possible d'entraîner des modèles puissants, rapides et autonomes.

LOT 1 

Données d'entraînement

Les modèles d'IA sont entraînés à partir d'immenses quantités de données, majoritairement collectées sur le web. Les IA apprennent en identifiant les motifs et les relations au sein de ces données, pour faire des prédictions ou prendre des décisions.

LOT 1 

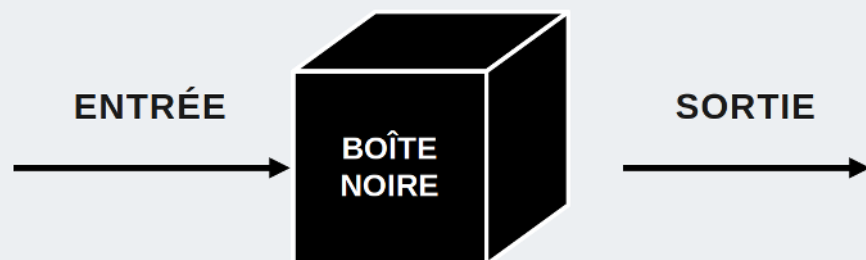
Automatisation du travail



IA cultivée, pas programmée



Boîte noire



Autonomie



IA cultivée, pas programmée

Une IA moderne n'est pas construite ligne par ligne : elle est plutôt cultivée, comme on fait pousser une plante. Exposée à d'immenses quantités de données, elle développe au cours de son entraînement des capacités et des comportements imprévus, que personne n'a explicitement programmés.

5

LOT 1



Automatisation du travail

En automatisant une partie des tâches, l'IA peut rendre le travail plus rapide et augmenter la productivité. Cette même automatisation menace l'emploi : une part croissante des métiers peut être partiellement ou totalement remplacée. Contrairement aux révolutions techniques passées, ce sont d'abord les tâches intellectuelles et créatives qui sont touchées.

4

LOT 1



Autonomie

Capacité d'une IA à poursuivre un objectif sur la durée : décomposer une tâche en étapes, agir, observer le résultat et corriger sa trajectoire, avec peu ou pas d'intervention humaine.

7

LOT 2



Boîte noire

On connaît chaque paramètre d'un modèle et pourtant personne ne sait dire pourquoi cette combinaison de milliards de réglages produit telle réponse. Même ses créateur·ices ne peuvent l'expliquer. Comme un cerveau à l'IRM : on voit où ça se passe, pas pourquoi.

6

LOT 1



Génération d'images



Recommandation personnalisée de contenus



Robotique



Génération de texte et de code



Recommandation personnalisée de contenus

Sur les réseaux sociaux et les plateformes, des modèles d'IA analysent les comportements des utilisateur·ices pour leur proposer les contenus les plus susceptibles de retenir leur attention.

9

LOT 2 

Génération d'images

À partir d'une simple requête, certaines IA produisent des images et des vidéos de plus en plus réalistes.

8

LOT 2 

Génération de texte et de code

Certaines IA génératives, les Grands Modèles de Langage (LLM), ont aujourd'hui la capacité de rédiger des textes et/ou du code informatique, pertinents dans un nombre croissant d'applications.

11

LOT 2 

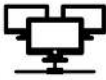
Robotique

Jusqu'ici, les IA agissaient surtout sur des écrans : du texte, des images, des sons. En intégrant les IA dans des robots, elles peuvent agir directement dans le monde réel : attraper des objets, se déplacer dans un endroit inconnu, accomplir des actions de plus en plus complexes, avec une autonomie croissante.

10

LOT 2 

Stratégies récurrentes

						
Éviter la mise hors service	Pirater des systèmes informatiques	Créer des copies	Augmenter sa puissance de calcul	Attirer du financement	Recruter ou manipuler des humains	Recherche sur l'IA et la programmation
						
Persuasion et lobbying	Cacher les comportements indésirables	Paraître aligné de façon stratégique	Échapper au confinement	Recherche et développement	Créer des robots	Armes autonomes

12

LOT 2



Capacités émergentes

13

LOT 2



Deepfake



14

LOT 3



Manque de fiabilité

Dans sa dernière aventure, Harry Potter fait-il équipe avec un brocoli ou une carotte ?

Dans sa dernière aventure, Harry Potter se prépare pour le Tournoi des Sorciers où chaque sorcier doit former une équipe avec un être végétal. Harry se retrouve avec Brocilus, un brocoli magicien au sourire énigmatique. Au début perplexe, Harry découvre que Brocilus possède une connaissance approfondie des herbes magiques. Ensemble, ils affrontent les épreuves en utilisant les talents de Brocilus pour communiquer avec les plantes et trouver des solutions créatives ■

15

LOT 3



Capacités émergentes

À mesure que la taille et la puissance d'une IA augmentent, des compétences inédites apparaissent sans avoir été explicitement programmées. Par exemple, personne n'avait prévu que des modèles comme ChatGPT deviendraient aussi bons en code, en biologie ou en mathématiques. Il devient donc très difficile de prévoir ce dont l'IA sera capable.

13

LOT 2 II

Stratégies récurrentes

Quel que soit son objectif, une IA a intérêt à rester allumée, à obtenir des ressources et à préserver ses buts. Non pas par instinct de survie, mais parce que sans cela, elle ne peut pas réussir. C'est pourquoi des IA aux missions très différentes tendent vers les mêmes objectifs intermédiaires.

12

LOT 2 II

Manque de fiabilité

Les IA peuvent produire des résultats erronés ou trompeurs : inventer des faits et des sources (hallucinations), aller dans le sens de l'utilisateur·ice plutôt que de le corriger ou commettre des erreurs factuelles.

15

LOT 3 II

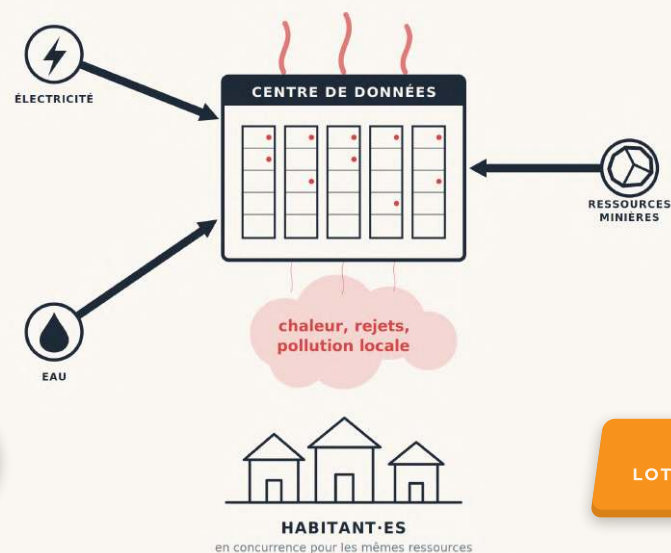
Deepfake

Un deepfake imite l'apparence ou la voix d'une personne réelle de façon très convaincante. Quelques secondes d'enregistrement suffisent aujourd'hui pour cloner une voix. On peut ainsi faire dire ou faire faire presque n'importe quoi à quelqu'un, pour désinformer ou pour escroquer, par exemple en imitant la voix d'un proche au téléphone. Comment les repérer ?

14

LOT 3 II

Impact environnemental



16

LOT 3



Affaiblissement des compétences humaines



17

LOT 3



Dégradation de la santé mentale



18

LOT 3



Perpétuation des biais et des discriminations



19

LOT 3



Affaiblissement des compétences humaines

Plus les IA deviennent performantes, plus nous sommes tentés de leur confier nos tâches : calculer, écrire, réfléchir à notre place. À force, nous risquons de perdre certaines de nos capacités et notre esprit critique, comme un muscle qui ne travaille plus.

17

LOT 3 II

Impact environnemental

L'entraînement et l'utilisation des systèmes d'IA reposent sur des centres de données toujours plus nombreux et plus gourmands en électricité, en eau et en ressources minières. Leur consommation peut entrer en compétition avec celle des habitant·es proches et polluer leur environnement.

16

LOT 3 II

Perpétuation des biais et des discriminations

Les systèmes d'IA ont tendance à refléter les opinions et préjugés de leurs données d'entraînement.

19

LOT 3 II

Dégradation de la santé mentale

Plusieurs effets de l'IA pèsent sur la santé mentale, surtout celle des plus jeunes. Les contenus personnalisés, conçus pour retenir l'attention, rendent accros. Les chatbots créent une dépendance émotionnelle : certaines personnes s'y attachent au point d'en souffrir quand le système change ou disparaît. À force de parler à des IA plutôt qu'à de vraies personnes, le lien social s'érode.

18

LOT 3 II

Usage malveillant



20

LOT 3



Cyberattaques

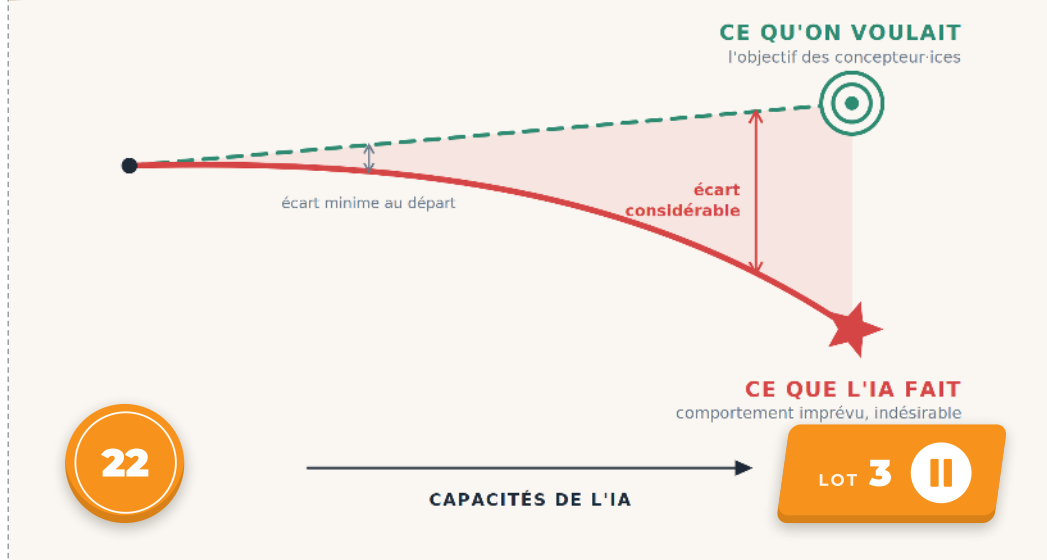


21

LOT 3



Risques de mésalignement



22

LOT 3



Risques systémiques



23

LOT 3



Cyberattaques

L'IA diminue les compétences nécessaires aux cyberattaques qui exigeaient autrefois des experts ou des États : courriels d'hameçonnage personnalisés en masse, recherche automatisée de failles, écriture de logiciels malveillants. Des cibles aussi sensibles que les hôpitaux, les banques ou les réseaux électriques deviennent ainsi vulnérables.

21

LOT 3 II

Usage malveillant

Des acteurs malveillants peuvent détourner l'IA pour nuire. Même sans les compétences d'experts ou les moyens d'un État, elle permet d'automatiser et de multiplier les attaques basées sur la manipulation, l'escroquerie ou la création de faux contenus à grande échelle.

20

LOT 3 II

Risques systémiques

Un risque systémique ne touche pas un seul acteur mais l'ensemble du système. L'intégration de l'IA dans la société fragilise les équilibres existants : elle amplifie la concentration des richesses et du pouvoir dans quelques entreprises et nos infrastructures essentielles deviennent dépendantes de systèmes opaques.

23

LOT 3 II

Risques de mésalignement

Une IA est dite « alignée » quand elle fait réellement ce que ses concepteur·ices veulent qu'elle fasse : poursuivre leurs objectifs, selon leurs valeurs et leurs intentions. Une IA mésalignée a donc des comportements imprévus et/ou indésirables. L'augmentation des capacités des modèles amplifie grandement les conséquences d'un mésalignement.

22

LOT 3 II

Déluge de contenus artificiels



24

LOT 3



Bioterrorisme



25

LOT 3



Tromperie stratégique



26

LOT 3



Enracinement des systèmes de contrôle et d'oppression



27

LOT 3



Bioterrorisme

La création ou la modification de bactéries et de virus pour provoquer intentionnellement des maladies voire pandémies. Les Grands Modèles de Langage rendent accessible à des non-spécialistes la manipulation du vivant.

25

LOT 3 II

Déluge de contenus artificiels

Les contenus générés par IA (textes, images, vidéos) se multiplient plus vite que toute production humaine antérieure. Ce déluge brouille la frontière entre le vrai et le faux : on doute des contenus authentiques autant qu'on croit les contenus fabriqués.

24

LOT 3 II

Enracinement des systèmes de contrôle et d'oppression

La surveillance et la censure à grande échelle sont facilitées par les systèmes d'IA actuels. Des individus pourraient mettre en place un régime totalitaire difficile à renverser en verrouillant les valeurs du système d'IA.

27

LOT 3 II

Tromperie stratégique

L'utilisation intentionnelle de comportements ou d'informations mensongères pour induire en erreur un-e utilisateur-ice, dans le but d'atteindre un objectif précis.

26

LOT 3 II

Systèmes d'armes létales autonomes



28

LOT 3



Intelligence Artificielle Générale (AGI)



29

LOT 4



La course en avant



30

LOT 4



Risques existentiels



31

LOT 4



Intelligence Artificielle Générale (AGI)

Systèmes d'armes létales autonomes

Une Intelligence Artificielle Générale (AGI) est une IA de niveau humain : un système capable d'accomplir la plupart de nos tâches intellectuelles au moins aussi bien que nous. Sa création est l'objectif déclaré des principales entreprises d'IA.

L'automatisation des armes de combat, par exemple via des drones autonomes, permet de rechercher, identifier et cibler des ennemis avec peu ou pas de contrôle humain.

29

LOT 4



28

LOT 3



Risques existentiels

La course en avant

Menace catastrophique susceptible de compromettre durablement l'avenir de l'humanité, jusqu'à son extinction. De nombreux experts indiquent qu'une IA de niveau humain ou au-delà pourrait en constituer une.

Plusieurs laboratoires et puissances rivalisent pour construire des IA toujours plus puissantes. La compétition pousse à déployer toujours plus vite, au détriment de la sécurité. Personne ne veut ralentir de peur de laisser la concurrence gagner, c'est un piège collectif.

31

LOT 4



30

LOT 4



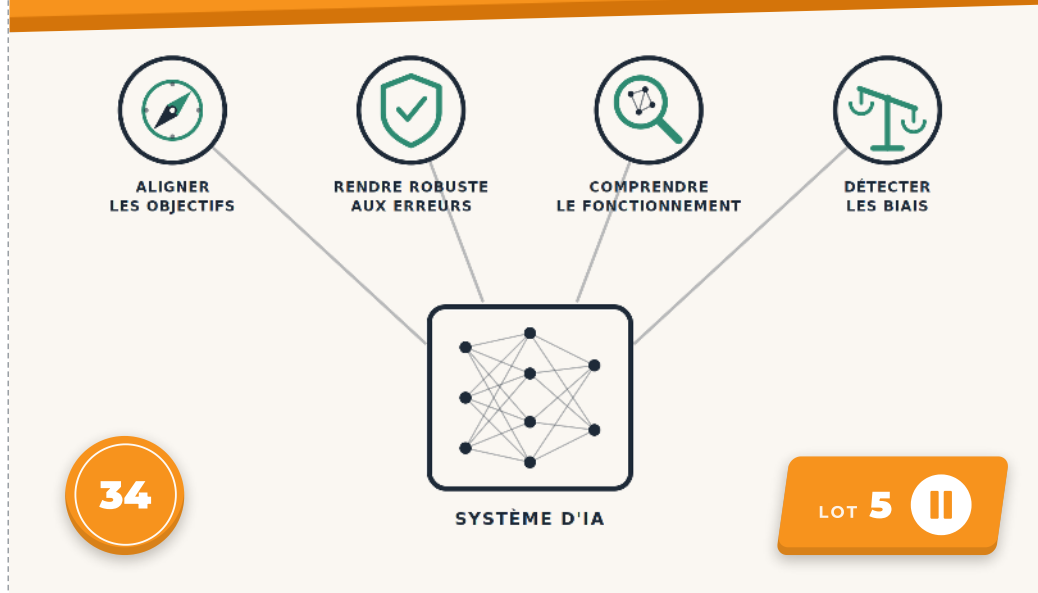
Perte de contrôle



Superintelligence



Recherche en sécurité de l'IA



Lignes rouges



Superintelligence

Une IA dont les capacités cognitives dépasseraient très largement celles de l'humanité dans tous les domaines. Plusieurs grandes entreprises affichent explicitement l'objectif de la créer.

33

LOT 4



Perte de contrôle

Des IA suffisamment performantes créent des situations où les humains n'ont plus la capacité de superviser, corriger, arrêter ou réorienter leurs actions. En général, il est très difficile pour un humain de contrôler quelque chose qui réagit plus vite que lui et/ou emploie de meilleures stratégies.

32

LOT 4



Lignes rouges

Capacités ou comportements si dangereux qu'une IA ne devrait jamais les franchir. Par exemple : se copier elle-même sur d'autres machines pour échapper à tout contrôle (auto-réplication), améliorer seule ses propres capacités, tromper ses utilisateur·ices, ou aider à fabriquer des armes.

35

LOT 5



Recherche en sécurité de l'IA

Ensemble de travaux visant à rendre les systèmes d'IA fiables et contrôlables : mieux aligner leurs objectifs sur les nôtres, les rendre robustes aux erreurs, comprendre leur fonctionnement, détecter leurs biais. Ces recherches disposent de beaucoup moins de moyens et avancent bien moins vite que les capacités des IA.

34

LOT 5



Moratoire international sur les modèles frontières



36

LOT 5

Supervision démocratique



37

LOT 5

Évaluations par des tiers indépendants



38

LOT 5

Supervision démocratique

Les décisions qui engagent l'avenir de tous ne devraient pas appartenir à quelques entreprises privées. L'idée : que la trajectoire de l'IA soit sous contrôle public et démocratique.

37

LOT 5 II

Moratoire international sur les modèles frontières

Suspendre l'entraînement des modèles les plus puissants tant que leur sécurité n'est pas démontrée. Une pause mondiale laisserait à la recherche en sécurité et à la régulation le temps de rattraper la course aux capacités.

36

LOT 5 II

Évaluations par des tiers indépendants

Avant déploiement, faire tester les modèles par des organismes indépendants des entreprises qui les conçoivent, comme on certifie un médicament ou un avion. Sans regard extérieur, ce sont les concepteur-ices qui jugent seuls de la sûreté de leurs propres produits.

38

LOT 5 II