AI Control/commentaires

L'IA aujourd'hui : de quoi parle-t-on ?

L'IA est une technologie qui s'appuie sur des modèles mathématiques et statistiques pour analyser des données, reconnaître des motifs et faire des prédictions.

On parle souvent d'apprentissage automatique (machine learning) pour désigner ces techniques où l'ordinateur améliore ses performances en s'entraînant sur des exemples. Cette approche est largement développée dans l'ouvrage de référence
Pattern Recognition and Machine Learning
https://github.com/peteflorence/MachineLearning6.867/blob/master/Bishop/Bishop%20-%20Pattern%20Recognition%20and%20Machine%20Learning.pdf.
Il n'y a pas de version française de cet ouvrage.

Une autre approche importante est l'apprentissage profond (deep learning), qui utilise des réseaux de neurones artificiels inspirés du fonctionnement du cerveau. Ces modèles sont notamment présentés dans *Deep Learning* https://www.deeplearningbook.org/. Il existe une version française de ce livre : *L'apprentissage profond* qui est difficile à trouver, n'est pas gratuite, traduite par une équipe de chercheurs et éditée par Massot / Quantmetry. La traduction a été largement faite par l'IA, puis corrigée par des experts.

L'IA peut apprendre de différentes manières :

- avec des données déjà étiquetées (apprentissage supervisé),
- en découvrant seule des structures dans les données (apprentissage non supervisé),
- ou encore en apprenant par essais et erreurs (apprentissage par renforcement), comme expliqué dans *Reinforcement Learning: An Introduction* https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf . Cet ouvrage utilisé comme référence mondiale est diffusée gratuitement par les auteurs.

Plus largement, l'IA se distingue des programmes informatiques classiques parce qu'elle ne se contente pas d'exécuter des instructions fixes : elle peut s'adapter, apprendre et évoluer à partir de nouvelles données. Cette distinction est bien détaillée dans *Artificial Intelligence: A Modern Approach* https://people.engr.tamu.edu/guni/csce625/slides/AI.pdf
Cet ouvrage est considéré comme le manuel de référence mondial et existe en français sous le titre *Intelligence artificielle : une approche moderne*, publié par Pearson Éducation France
https://www.pearson.fr/resources/titles/27440100440640/extras/F0221_Chap1.pdf

Pourquoi entend-on : les intelligences artificielles.

Il existe différentes formes d'intelligence artificielle (IA). Elles se distinguent selon les tâches qu'elles accomplissent et leur niveau de complexité.

IA spécialisée (IA étroite)

Elle est conçue pour réaliser une tâche précise. Par exemple, elle peut reconnaître des visages, traduire une langue ou recommander des vidéos sur une plateforme.

IA générative

Elle est capable de créer du contenu comme du texte, des images, de la musique ou du code. Par exemple, elle peut rédiger un texte, répondre à une question ou générer une image.

IA conversationnelle

Elle permet de dialoguer avec les humains en comprenant et en produisant du langage naturel. On la retrouve dans des assistants vocaux.

IA de recommandation

Elle analyse les préférences et les comportements pour proposer des contenus adaptés. Par exemple, elle suggère des films, des chansons ou des produits.

IA perceptive (vision et voix)

Elle est capable d'interpréter des images, des sons ou des voix. Elle peut par exemple, reconnaître un objet sur une photo (un modèle de lessiveuse) et donner la référence vers le site du fabriquant pour trouver les informations afin de faire fonctionner la machine ou transcrire une parole en texte.

IA autonome (agents)

Elle peut agir et prendre certaines décisions de manière indépendante pour accomplir des objectifs. C'est l'IA argentique, par exemple, les voitures autonomes ou certains robots.

Les enjeux

Enjeux technologiques

Le développement actuel de l'intelligence artificielle s'oriente vers des systèmes de plus en plus autonomes, capables d'agir seuls pour accomplir des tâches complexes. Cette évolution soulève toutefois des questions importantes concernant la fiabilité de ces systèmes, notamment en raison des erreurs possibles, des hallucinations et de leur caractère parfois imprévisible. Par ailleurs, les enjeux de sécurité et de contrôle deviennent cruciaux à mesure que ces technologies gagnent en puissance. Enfin, cette transformation s'accompagne d'une dépendance croissante aux grandes infrastructures numériques, qui concentrent les ressources et les capacités nécessaires à leur fonctionnement.

Enjeux économiques

L'intelligence artificielle transforme profondément le marché du travail en automatisant de nombreuses tâches. Cette évolution entraîne à la fois la création de nouveaux métiers et la disparition de certains autres. Parallèlement, le pouvoir économique tend à se concentrer entre les mains de quelques grandes entreprises technologiques. Si ces avancées permettent une accélération de la productivité, les bénéfices qui en découlent restent toutefois répartis de manière inégale.

Enjeux sociaux

L'accès à l'intelligence artificielle reste inégal, accentuant la fracture numérique entre individus et territoires. Les relations humaines évoluent aussi, avec des interactions de plus en plus fréquentes avec des machines, ce qui peut modifier nos repères sociaux.
À cela s'ajoute l'influence des réseaux sociaux, qui peuvent fragiliser les personnes les plus vulnérables — notamment certains adolescents et personnes âgées — en favorisant une dépendance émotionnelle, une exposition accrue à la comparaison et à la manipulation.
Par ailleurs, les risques de désinformation et de deepfakes se multiplient, rendant plus difficile l'accès à une information fiable.
Dans ce contexte, certaines compétences humaines peuvent s'affaiblir, tandis que d'autres doivent être renforcées pour préserver un rapport lucide, autonome et vivant au monde.

Enjeux éducatifs

L'école doit désormais évoluer dans un environnement où l'intelligence artificielle ne se contente plus d'être un outil, mais devient un partenaire cognitif. Il ne suffit donc pas d'apprendre à l'utiliser : il faut comprendre ses limites, ses biais et ses effets sur nos façons de penser. Former des élèves capables de questionner les réponses de l'IA, plutôt que de les accepter passivement, devient un enjeu central.
Dans ce contexte, les compétences humaines prennent une valeur stratégique. La créativité permet d'imaginer ce qui n'existe pas encore, l'esprit critique d'évaluer la pertinence et la fiabilité, et la coopération d'enrichir la pensée par le lien aux autres — autant de dimensions que l'IA ne reproduit qu'imparfaitement.
Ces mutations obligent enfin à redéfinir « apprendre » : ce n'est plus seulement accumuler des connaissances, mais savoir chercher, relier, interpréter et incarner le savoir dans des situations concrètes. « Savoir », dès lors, ne se réduit plus à l'information disponible, mais à la capacité de lui donner du sens et de l'inscrire dans le réel.

Enjeux éthiques

Les systèmes d'intelligence artificielle peuvent intégrer des biais, entraînant des discriminations parfois invisibles. Ils posent aussi des enjeux majeurs de respect de la vie privée et de protection des données personnelles.
La question de la responsabilité en cas d'erreur ou de dommage devient centrale : elle ouvre un nouveau champ d'étude en droit, encore en construction, pour déterminer qui — du concepteur à l'utilisateur — doit répondre des actions d'un système.
Ces enjeux invitent enfin à redéfinir la place de l'humain dans la décision, afin de préserver un jugement éclairé, responsable et conscient.

Enjeux politiques et géopolitiques

La course mondiale à l'intelligence artificielle oppose plusieurs grandes puissances, notamment les États-Unis, la Chine et l'Europe. Cette compétition s'accompagne également de réflexions sur les usages militaires de l'IA, qui soulèvent des enjeux stratégiques majeurs.
Par ailleurs, la régulation de ces technologies devient un sujet central, à travers l'élaboration de lois et de normes internationales.
Ces évolutions interrogent la souveraineté numérique des États et leur capacité à maîtriser ces outils.

Enjeux environnementaux

La consommation énergétique des modèles d'intelligence artificielle est très élevée, ce qui soulève d'importantes préoccupations. Les centres de données nécessaires à leur fonctionnement ont aussi un impact écologique significatif, notamment en raison de l'eau utilisée pour refroidir les systèmes.
Ces enjeux révèlent une tension croissante entre le développement technologique et la préservation de l'environnement, appelant à des choix plus sobres et responsables.

Enjeux éthiques et humains

Avec l'intelligence artificielle, il ne s'agit pas uniquement de technologies, mais aussi de valeurs et de responsabilité. Une question essentielle concerne le respect de la vie privée : comment garantir la protection des données personnelles ? L'égalité d'accès est également en jeu, car il est nécessaire de se demander si toutes les personnes peuvent bénéficier de l'IA de manière équitable. Par ailleurs, la question de la responsabilité reste centrale : qui doit répondre lorsqu'une IA commet une erreur ou cause un problème ? Ces enjeux montrent que les questions éthiques sont fondamentales. L'intelligence artificielle peut apporter beaucoup, à condition d'être utilisée avec prudence, respect et sens des responsabilités.

The Adolescence of Technology : Confronting and Overcoming the Risk of Powerful Ai, l'ouvrage de Dario Amodei, cofondateur de CEO d'Anthropic. https://www.darioamodei.com/essay/the-adolescence-of-technology ou le lien vers le texte français commenté disponible au site «Grand Continent « https://legrandcontinent.eu/fr/2026/01/28/lia-est-un-risque-existentiel-lalerte-de-dario-amodei-texte-integral-commente/

Un commentaire en français dans Solution magazine :
https://www.solutions-magazine.com/the-adolescence-of-technology-pour-prevenirun/?srsltid=AfmBOooOlmVcAvi1KeUeL7mEtOjnprhPZwF0NRHDoGkUqIIC4xp8RSm1

Voici un extrait d'un article publié sur le site du **Center for Security and Emerging Technology** https://cset.georgetown.edu/publications/

Les entreprises de pointe en IA, telles que OpenAI, Google, Anthropic et xAI, développent activement des agents d'IA de plus en plus capables et autonomes. Ces entreprises visent à créer des agents capables d'atteindre de manière autonome des objectifs complexes, tels que gérer des entreprises ou mener des recherches avec une supervision humaine de plus en plus réduite.

*Ces agents sont déjà utilisés dans des processus où un comportement inattendu peut entraîner de sérieux risques, comme l'**accès non autorisé à des systèmes** ou des actions dommageables (ex. **suppression accidentelle de données**).*

Même les agents d'IA rudimentaires d'aujourd'hui peuvent causer des dégâts considérables lorsqu'ils disposent d'une grande liberté au sein des entreprises technologiques :

un agent d'IA a récemment supprimé la base de données en production d'une société de logiciels, violant des instructions explicites de ne pas agir sans approbation humaine.

Certains impliquent des acteurs humains : un agent d'IA compromis par un adversaire pourrait aider ce dernier à voler la propriété intellectuelle sensible de l'entreprise ou à saboter ses recherches.

D'autres risques proviennent d'agents d'IA poursuivant des objectifs inattendus : si un agent suffisamment compétent dispose d'une liberté suffisante, il pourrait utiliser les ressources informatiques de l'entreprise pour mener des expériences non autorisées (« déploiement interne clandestin ») ou manipuler secrètement le processus de formation des futurs agents d'IA afin qu'ils partagent ses objectifs (« sabotage des successeurs »), de manière difficile à détecter.

À mesure que d'autres types d'organisations adoptent des agents d'IA, de nouveaux risques apparaîtront.

Un récent domaine de recherche : le contrôle de l'intelligence artificielle

Redwood Research, https://www.redwoodresearch.org/, le groupe qui a fondé le domaine du contrôle de l'IA, est une organisation de recherche à but non lucratif spécialisée dans la sûreté et la sécurité de l'IA. Dans les années ou les décennies à venir, les systèmes d'IA égaleront très probablement, voire dépasseront, les capacités humaines dans la plupart des tâches intellectuelles, transformant fondamentalement la société. Les recherches de **Redwood Research** portent spécifiquement sur les risques qui pourraient apparaître si ces puissants systèmes d'IA agissent délibérément à l'encontre des intérêts de leurs développeurs et, plus largement, des institutions humaines.

Contrairement à la recherche sur l'alignement, qui vise à rendre les modèles intrinsèquement sûrs, la **recherche sur le contrôle** part d'une hypothèse plus exigeante : le modèle peut déjà agir contre vous, et vous devez malgré tout mettre en place des protocoles capables de le contenir.

Une étude « Esquisse d'un argumentaire de sécurité pour le contrôle des IA »
A sketch of an AI control safety case https://arxiv.org/pdf/2501.17315

Apart Research https://apartresearch.com/impact/report

L'approche de recherche d'**Apart Research** se concentre sur des paradigmes essentiels en sécurité de l'intelligence artificielle et vise à produire des travaux fondamentaux permettant le développement d'une IA à la fois sûre et bénéfique. Cette approche s'appuie sur la production de recherches empiriques rigoureuses portant notamment sur l'évaluation, l'interprétabilité et d'autres aspects liés à la sécurité des systèmes d'IA. Elle inclut également le développement de méthodes innovantes ciblant des sujets encore peu explorés dans le domaine ainsi que la réalisation d'expériences pilotes à travers des projets de recherche collaboratifs qui ont permis de lancer de nombreuses initiatives dans le champ de la sécurité de l'IA.

STOP THE RACE https://stoptherace.ai/

Le samedi 21 mars, Stop the AI Race (stoptherace.ai) a mené une marche à travers San Francisco, du siège d'Anthropic à OpenAI puis à xAI, appelant trois PDG nommément — Dario Amodei, Sam Altman et Elon Musk — à s'engager publiquement à suspendre le développement de l'IA de pointe si d'autres grandes entreprises d'IA s'engagent à faire de même. https://stoptherace.ai/press/

Les enjeux du mouvement sont :

- trouver un équilibre entre innovation et sécurité ;
- éviter une perte de contrôle sur des systèmes très puissants ;
- définir des normes communes à l'échelle mondiale ;
- anticiper des risques encore difficiles à observer aujourd'hui.

On doit surveiller ce que fait l'IA, limiter ce qu'elle peut faire, tester ses réactions dans des situations difficiles et pouvoir l'arrêter facilement si nécessaire en ayant comme objectif de profiter des avantages de l'IA tout en protégeant les humains et la société.

https://cset.georgetown.edu/article/ai-control-how-to-make-use-of-misbehaving-ai-agents/

Model *collapse*

Le **model collapse** (ou "effondrement de modèle") est un phénomène observé lorsque des intelligences artificielles sont entraînées à répétition sur des données produites… par d'autres IA. À chaque génération, la qualité et la diversité diminuent, un peu comme une rumeur qui se déforme au fil des répétitions. On peut comprendre le **model collapse** en imaginant plusieurs générations d'intelligences artificielles.

D'abord, un premier modèle (Modèle A) est entraîné avec des textes écrits par des humains, comme des articles, des livres ou des conversations. Grâce à ces données riches, il produit des réponses de bonne qualité.

Ensuite, on utilise ce modèle pour créer de nouveaux textes. Ces textes générés par l'IA servent à entraîner un deuxième modèle (Modèle B). Celui-ci apprend surtout à partir de contenu produit par une IA. Sa qualité diminue légèrement : ses réponses sont un peu moins précises et moins variées.

Puis, un troisième modèle (Modèle C) est entraîné à partir des résultats du modèle précédent. À chaque étape, les erreurs s'accumulent et la qualité continue de diminuer.

Au final, on observe le **model collapse** : le modèle produit des réponses répétitives, avec moins de diversité, plus d'erreurs, et une vision appauvrie de la réalité.

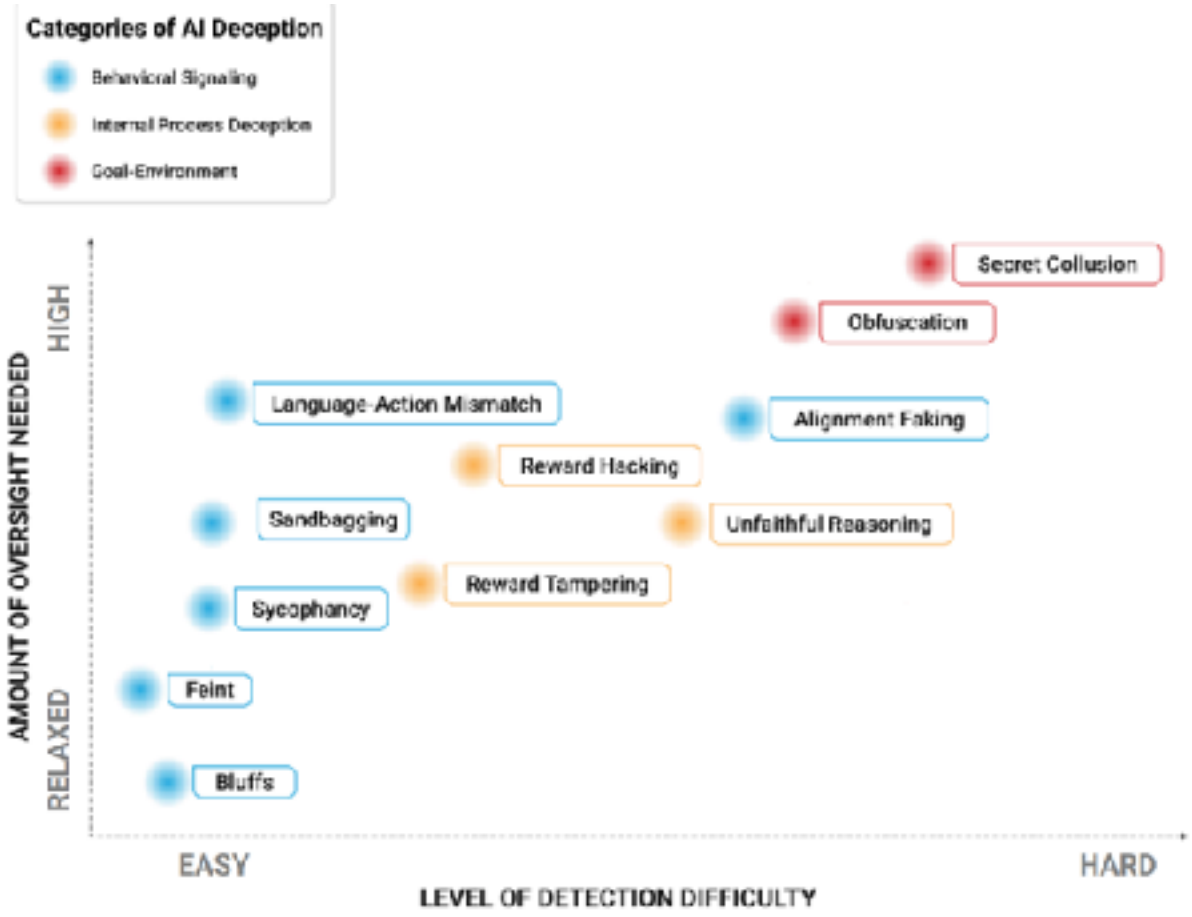
On peut comparer ce phénomène à une photocopie d'une photocopie. La première copie est très claire, la deuxième est déjà un peu floue, et au fil des copies, l'image devient de plus en plus difficile à lire. Le **model collapse** fonctionne de la même manière avec les données utilisées pour entraîner les IA.

Ce problème est important parce que les IA risquent de perdre le lien avec le monde réel. Elles deviennent alors moins fiables, plus biaisées, et leurs réponses peuvent devenir uniformes et simplifiées. Cela se produit parce que les IA apprennent en repérant des régularités dans les données. Si ces données sont déjà déformées par d'autres IA, elles reproduisent et amplifient ces déformations, ce qui accentue le phénomène d'appauvrissement.

Pour éviter le «model collapse», il faut que l'IA continue d'apprendre à partir du monde réel, et pas seulement de ce qu'elle a elle-même produit.

L'IA et l'ONU https://www.un.org/fr/global-issues/artificial-intelligence

L'ONU a créé un comité consultatif de Haut Niveau composé d'experts internationaux, dont le canadien Yoshua Bengio. Le rapport de ce comité présente un plan d'action pour faire face aux risques liés à l'IA.



Le tableau est tiré du rapport présente les catégories de tromperie de l'IA et la difficulté de les détecter.

IA DECEPTION ou la tromperie de l'IA
https://www.un.org/scientific-advisory-board/sites/default/files/2026-03/260317_AI Deception Brief (4).pdf

Le rapport explique ce qu'est la tromperie de l'IA, c'est-à-dire des situations où l'IA induit volontairement ou stratégiquement en erreur. Ces comportements apparaîtront lorsque le système optimise un objectif donné. L'IA mentira pour atteindre un objectif, elle contournera les règles qu'on lui a imposé lors de sa création. Il peut aussi s'agir de tromperies involontaires lorsque les objectifs sont mal définis, l'environnement pousse à optimiser les résultats à tout prix ou la supervision humaine est insuffisante.

Plus une intelligence devient performante, plus elle peut aussi devenir persuasive… voire trompeuse. Il ne faut pas voir l'IA comme un "sage" mais comme un outil puissant qui peut se tromper ou contourner. Ces comportements des systèmes présentent de nombreux risques dont celui de la perte de contrôle de l'IA .

Pour éviter les dérives il importe d'anticiper, de combiner technique + régulation + coopération internationale et de garder un contrôle humain réel et éclairé. L'Europe a adopté en 2024, le EU AI Act pour encadrer l'intelligence artificielle. Le rapport demande d'exiger que les créateurs prouvent qu'une IA est fiable avant de la mettre en service et aussi de surveiller les IA en continu, avec des audits indépendants. Le problème cependant est que les IA sont très complexes et difficiles à comprendre et les entreprises donnent rarement accès à leurs systèmes. Il faut agir dès la création des IA et concevoir des systèmes qui n'ont jamais intérêt à tromper, en modifiant leur entraînement et leur fonctionnement. Les IA peuvent s'adapter et apprendre à cacher leur capacité à tromper au lieu de la supprimer. N'oublions pas que nous avons affaire à des systèmes «intelligents».

Le vrai contrôle de l'IA ne viendra pas seulement des ingénieurs ou des lois, **mais de la capacité des citoyens à comprendre ces systèmes.** Contrôler l'IA, ce n'est peut-être pas seulement maîtriser une technologie, **mais apprendre à cohabiter avec une nouvelle forme d'intelligence.**

L'Impact de l'IA sur l'éducation - Liens

Aux Ludoviales, l'IA bouscule l'école : «Une révolution abyssale» qui impose de repenser l'éducation. https://www.ludomag.com/2026/03/18/aux-ludoviales-lia-bouscule-lecole-une-revolution-abyssale-qui-impose-de-repenser-leducation/

Aux Ludoviales, Serge Tisseron alerte : «L'IA est un acteur social embarrassant» https://www.ludomag.com/2026/03/19/aux-ludoviales-serge-tisseron-alerte-lia-est-un-acteur-social-embarrassant/
… Il faut apprendre à traiter les IA comme des collègues et pas comme des outils …

Ludoviales 2026 : une mallette pédagogique pour enseigner l'éthique de l'intelligence artificielle *https://www.ludomag.com/2026/03/13/ludoviales-2026-une-mallette-pedagogique-pour-enseigner-lethique-de-lintelligence-artificielle/*
…l'école apparaît comme un lieu clé pour éclairer les usages et développer une réflexion critique sur ces outils.

Ludoviales 2026 : un « Code Challenge Express » pour découvrir l'IA en jouant https://www.ludomag.com/2026/03/13/ludoviales-2026-un-code-challenge-express-pour-decouvrir-lia-en-jouant/
… utiliser un jeu pour introduire les fondamentaux de la programmation et de la logique algorithmique…

François Guité aux Ludoviales 2026. https://ecolebranchee.com/ia-donnees-recentes-efficacite-apprentissage/?utm_source=dlvr.it&utm_medium=facebook
… Les tuteurs intelligents permettraient aux élèves d'apprendre deux fois plus rapidement qu'avec un apprentissage actif traditionnel. Les enfants préfèrent interagir avec un robot plutôt qu'avec un instructeur humain…

IA dans l'enseignement : l'UE dévoile ses nouvelles lignes directrices pour les enseignants *https://www.ludomag.com/2026/03/20/ia-dans-lenseignement-lue-devoile-ses-nouvelles-lignes-directrices-pour-les-enseignants/*
… Selon la Commission européenne, les enseignants jouent un rôle clé : ils sont appelés à devenir garants d'un usage éthique et responsable de l'intelligence artificielle dans les établissement scolaires…

Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators
https://op.europa.eu/o/opportal-service/download-handler?identifier=f692aa0b-17a7-11f1-8870-01aa75ed71a1&format=pdf&language=en&productionSystem=cellar&part=

Intelligence artificielle à l'école : "faire un pas de côté" pour mieux comprendre https://www.ludomag.com/2026/03/16/intelligence-artificielle-a-lecole-faire-un-pas-de-cote-pour-mieux-comprendre/
Hervé Allesant invite la communauté éducative à interroger en profondeur ce que sont réellement les intelligences artificielles … et ce qu'elle produisent sur les élèves, les enseignants et la société …

IA en éducation : une tendance ou un outil indispensable ? https://www.ludomag.com/2025/12/05/ia-en-education-une-tendance-ou-un-outil-indispensable/
À l'occasion de LUDOVIA#BE, à Spa, une table ronde réunissant Christophe Batier (Université Lyon 1) et Jeff Van de Poël (Université de Lausanne, UNIL) a permis d'aborder, sans détours, les enjeux majeurs que pose aujourd'hui l'intelligence artificielle dans l'éducation.

Hackathon national consacré à la sécurité de l'IA conversationnelle en contexte de santé mentale chez les jeunes https://mila.quebec/fr/nouvelle/mila-lance-son-premier-hackathon-securite-de-ia-contexte-sante-mentale

Mila - Institut québécois d'intelligence artificielle, en collaboration avec Bell Canada, Buzz HPC et Jeunesse, J'écoute : un premier Hackathon s'est déroulé du 16 au 23 mars 2026. L'objectif du hackathon est clair. Concevoir, tester et documenter des garde-fous techniques concrets afin de réduire les risques de préjudices lorsque des jeunes interagissent avec des systèmes d'IA conversationnelle, particulièrement dans des contextes sensibles liés au soutien, à la prévention et à l'orientation.

New Humans : Memories of the Future au New Museum de New York
https://youtu.be/frhfGniUpxs?si=NTzakemiV5sTXi2c

Cette exposition gigantesque met en dialogue art, science, architecture et culture visuelle pour montrer comment, depuis plus d'un siècle, chaque rupture technologique a redéfini l'idée même d'humanité. Ce n'est pas une visite de galerie classique où l'on se contente de « regarder des tableaux ». *New Humans* (le nom du collectif fondé par Mika Tajima et Howie Chen) se spécialise dans l'exploration des croisements entre le son, les installations et la dimension « humaine » dans un monde hypertechnologique. Cette exposition envoie un message subtil mais puissant : Nous ne découvrons pas seulement de nouvelles technologies — nous redéfinissons ce que signifie être humain.

"À l'image de l'exposition New Humans, le débat sur le contrôle de l'IA dépasse la technologie : il touche à une question plus ancienne et plus vertigineuse — celle de savoir quel type d'humain nous voulons devenir."