Intelligence artificielle:

définition, impact sociétal, domaines, bénéfices et risques, future réglementation

Journée de la recherche

Marianne Verdier, Professeur d'économie, Université Paris Panthéon-Assas,
CRED, Centre de Recherches en Economie et Droit,
Responsable de la Chaire Finance Digitale et du Master 2 Finance

PANTHÉON-ASSAS UNIVERSITÉ PARIS

CRED
Centre de recherche en économie et droit

Définition de l’intelligence artificielle

- **Selon une définition de l’OCDE:**
  - *`An AI system is a **machine-based system** that can, for a given set of **human** defined **objectives**, make **predictions**, **recommendations**, or **decisions** influencing **real or virtual** environments'.*
- Cette définition est similaire à celle donnée par la Commission Européenne dans la réglementation du 13 juin 2004 (AI Act).

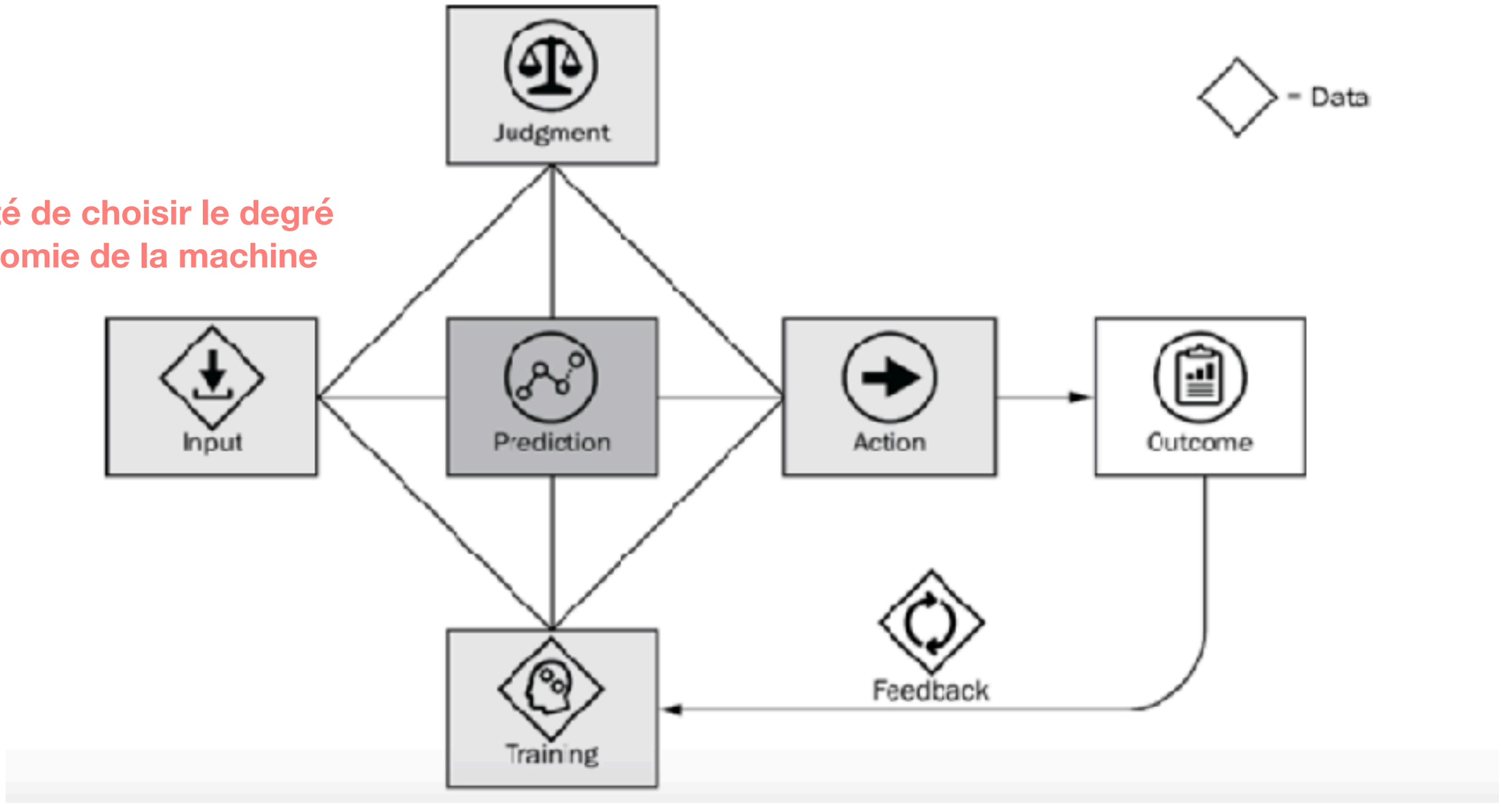
- **Plusieurs points importants distinguent l’intelligence artificielle d’autres technologies digitales:**
  - 1) la possibilité de **laisser une machine apprendre** avec différents degrés d’autonomie, pouvant impliquer l’humain à différents niveaux
  - 2) l’usage de très larges sources de données qui peuvent être combinées pour générer de nouvelles données et de nouvelles informations
    - Une prédiction correspond à la génération d’une **nouvelle information** à partir d’anciennes informations
  - 3) la possibilité dans certains cas de **laisser une machine décider** de façon autonome

Définition de l’intelligence artificielle

- Le règlement IA du 13 juin 2024 établit une distinction entre:
  - 1) **<u>les systèmes d’IA (art. 3, 1)):</u>**
    - systèmes automatisés conçus pour fonctionner à différents niveaux d’autonomie (…) pour des objectifs explicites ou implicites (…)
  - 2) **<u>les modèles d’IA à usage général (art. 3, 63)):</u>**
    - modèle entraîné avec un grand nombre de données utilisant l’auto-supervision à grande échelle (…) pouvant exécuter un certain nombre de tâches distinctes (…) et pouvant être intégré dans une variété d’applications en aval (…)

Définition de l'intelligence artificielle

Possibilité de choisir le degré d'autonomie de la machine

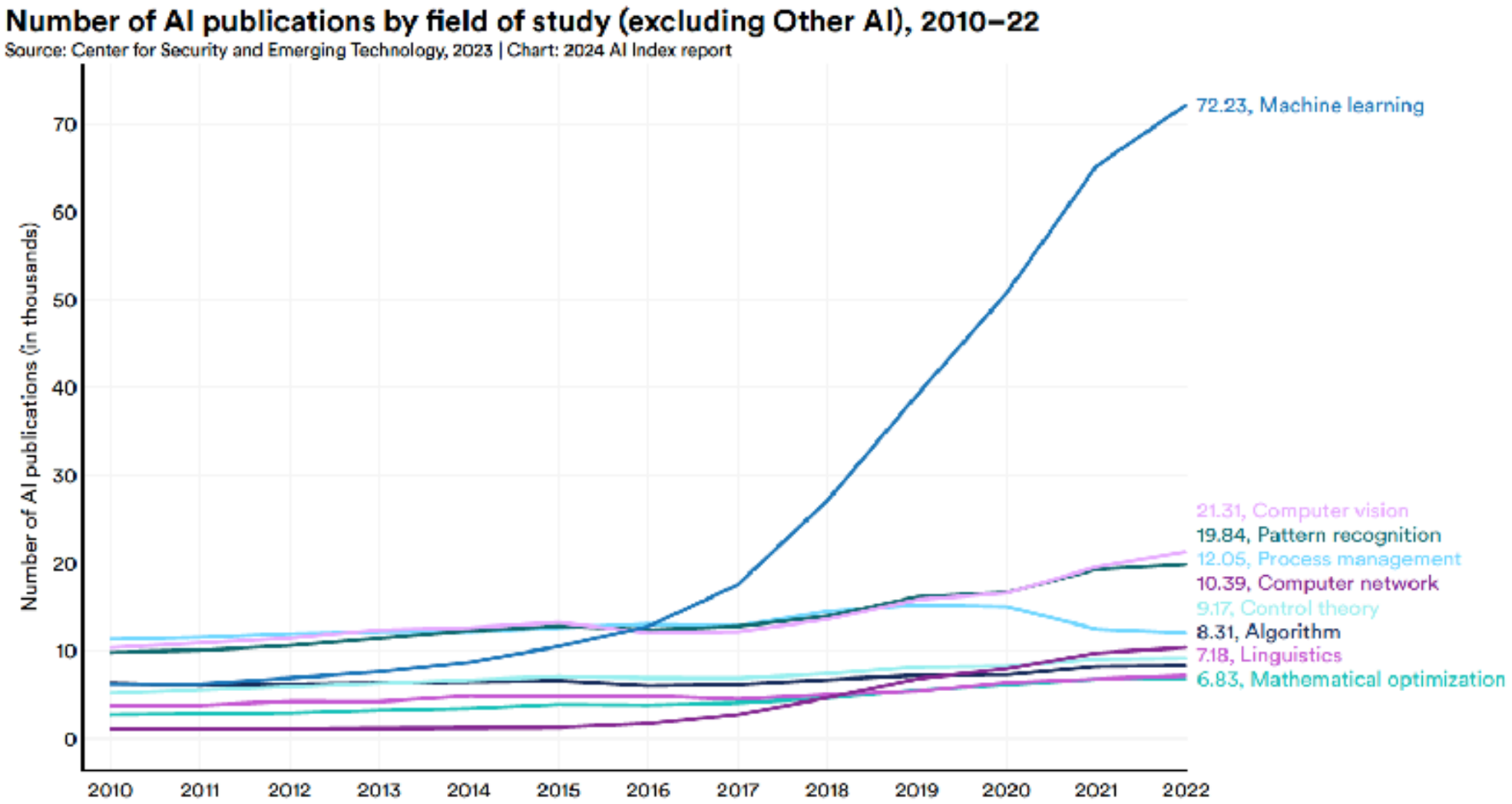


Source: Gans (2024), presentation at the Center for Competition Economics https://cceeco.org/events

Les grands domaines de l’intelligence artificielle

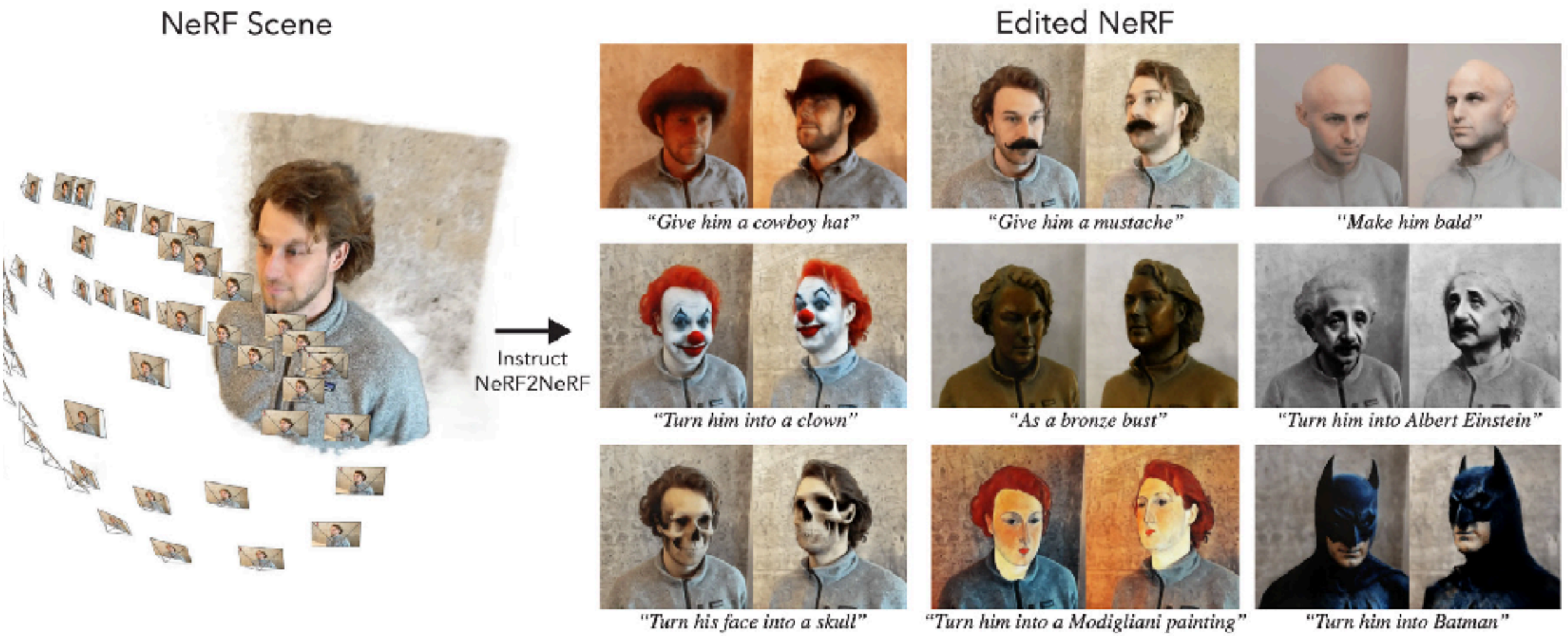
- **Robotic Process Automation (RPA):** robots logiciels pour exécuter des tâches répétitives auparavant effectuées par des humains.
    - => Usage pour réduire des tâches administratives.
- **Machine Learning (ML):** programme capable de trouver un modèle ou de réaliser des prédictions à partir de données.
    - => Usage pour réduire des faux positifs.
- **Deep Learning:** une sous-catégorie du ML qui utilise des réseaux neuronaux multicouches, appelés réseaux neuronaux profonds, pour simuler le pouvoir de décision complexe du cerveau humain.
    - => Usage par exemple dans les voitures autonomes.
- **IA générative:** intelligence artificielle capable de générer de nouveaux contenus comme du texte ou des images en réponse à une requête.

Les grands domaines de l’intelligence artificielle



Source: Artificial Intelligence Index report, Stanford University, 2024

Exemple: produire de nouvelles informations



Exemple: image+langage -> nouvelle image
Modèle Instruct-NeRF2NeRF développé par des chercheurs de l'université de Berkley pour générer de nouvelles images à partir de texte

Source: Artificial Intelligence Index report, Stanford University, 2024

Exemples: produire de nouvelles informations

- Reconnaître et générer du texte à partir de texte et de données de langage (LLM, large langage models).

- Combiner des données pour générer des indicateurs d'aide à la décision (exemple: décisions d'octroi de crédit, détection de la fraude, décisions médicales).

- Identifier et générer des images ou des vidéos à partir de texte.

High-quality generation of milk dripping into a cup of coffee

Source: Blattmann et al., 2023

Exemple: la détection de la criminalité financière

- Le 19 nov. 2020, Nasdaq.inc a acheté l'entreprise Verafin pour 2,75 milliards de dollars.

- Cette entreprise a investi dans des solutions d'intelligence artificielle pour prédire la fraude et valider l'identité des consommateurs bancaires.

« Verafin's capabilities will be available to the global network of nearly 250 banks, exchanges, broker-dealers and buy-side organizations, and regulatory authorities that rely on Nasdaq's technology to detect market manipulation and abuse today. »

Exemple: la détection de la criminalité financière

Source: Verafin, Artificial Intelligence, outsmarting financial crimes

L’organisation industrielle du secteur (1)

- ***Qui développe les modèles? Existe-t-il des barrières à l’entrée?***
  - -> rôle prépondérant des big techs pour les modèles d’IA (Open AI détenu à 49% par Microsoft Azur, Google, Meta), mais pas uniquement (ex de Verafin). Barrières fortes: puissance de calcul, cloud, volume de données, expérience.
  - -> le règlement IA européen ne s’applique pas uniquement aux fournisseurs, mais aussi aux déployeurs (utilisateurs), importateurs et distributeurs de systèmes d’IA, aux fabricants de produits remplissant certaines conditions, aux mandataires des fournisseurs…

- ***Est-ce que les modèles sont développés pour plusieurs industries ou pour des applications spécifiques?***
  - -> rôle des modèles de fondation (à usage général) versus systèmes développés pour un objectif précis

L’organisation industrielle du secteur (2)

- ***Qui peut avoir accès aux données d’entraînement? Comment les données sont-elles choisies, protégées?***

  - -> cela dépend, question de l’ouverture de l’accès aux modèles (open-source) et du respect de la vie privée

- ***Quels sont les coûts et les bénéfices de ces modèles?***

  - -> recherches pour évaluer les coûts de mise en oeuvre (puissance de calcul, empreinte carbone, collecte de données), les coûts des risques associés aux résultats, les coûts de long terme liés à l’adoption de la technologie

  - -> question de l’évaluation de la performance des algorithmes par rapport aux humains

Les risques liés au développement de l'intelligence artificielle

- **Les risques relatifs au marché du travail**
    - Question de l'automatisation des tâches et de la complémentarité entre travail humain et technologie (Acemoglu, 2024)
- **Les risques relatifs à l'absence de maîtrise de la technologie**
    - Les risques de gouvernance concernant les données qui sont de nature « sociale » et génèrent des externalités
    - Les risques relatifs à l'usage de l'intelligence artificielle dans les domaines pour lesquels des considérations éthiques ou morales sont importantes (exemple: santé)
    - Les risques relatifs au manque de transparence, d'explicabilité, à la présence de biais
    - Les risques relatifs à l'exercice du pouvoir de marché
- **Les risques de transformation des sociétés démocratiques et du fonctionnement du cerveau humain pour prendre des décisions**
    - Erosion de la capacité démocratique des sociétés, confusion de l'information, manipulations de groupes de personnes

Les risques liés au développement de l'intelligence artificielle

Responsible AI dimension	Definition	Example
Data governance	Establishment of policies, procedures, and standards to ensure the quality, security, and ethical use of data, which is crucial for accurate, fair, and responsible AI operations, particularly with sensitive or personally identifiable information.	Policies and procedures are in place to maintain data quality and security, with a particular focus on ethical use and consent, especially for sensitive health information.
Explainability	The capacity to comprehend and articulate the rationale behind AI decisions, emphasizing the importance of AI being not only transparent but also understandable to users and stakeholders.	The platform can articulate the rationale behind its treatment recommendations, making these insights understandable to doctors and patients, ensuring trust in its decisions.
Fairness	Creating algorithms that are equitable, avoiding bias or discrimination, and considering the diverse needs and circumstances of all stakeholders, thereby aligning with broader societal standards of equity.	The platform is designed to avoid bias in treatment recommendations, ensuring that patients from all demographics receive equitable care.
Privacy	An individual's right to confidentiality, anonymity, and protection of their personal data, including the right to consent and be informed about data usage, coupled with an organization's responsibility to safeguard these rights when handling personal data.	Patient data is handled with strict confidentiality, ensuring anonymity and protection. Patients consent to whether and how their data is used to train a treatment recommendation system.
Security and safety	The integrity of AI systems against threats, minimizing harms from misuse, and addressing inherent safety risks like reliability concerns and the potential dangers of advanced AI systems.	Measures are implemented to protect against cyber threats and ensure the system's reliability, minimizing risks from misuse or inherent system errors, thus safeguarding patient health and data.
Transparency	Open sharing of development choices, including data sources and algorithmic decisions, as well as how AI systems are deployed, monitored, and managed, covering both the creation and operational phases.	The development choices, including data sources and algorithmic design decisions, are openly shared. How the system is deployed and monitored is clear to healthcare providers and regulatory bodies.

Figure 3.1.1

Source: Artificial Intelligence Index report, Stanford University, 2024

Les risques liés au développement de l'intelligence artificielle

Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs

Yuxia Wang[1,2*] **Haonan Li**[1,2*] **Xudong Han**[1,2*]
Preslav Nakov[2] **Timothy Baldwin**[2,3]
[1]LibrAI [2]MBZUAI
[3]The University of Melbourne
{yuxia.wang, haonan.li, xudong.han}@mbzuai.ac.ae

Exemple de travail de recherche sur le recensement des risques des modèles d'intelligence artificielle et sur la comparaison des LLMs (2023)

Les risques liés au développement de l'intelligence artificielle

Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs

Yuxia Wang[1,2*] **Haonan Li**[1,2*] **Xudong Han**[1,2*]
Preslav Nakov[2] **Timothy Baldwin**[2,3]
[1]LibrAI [2]MBZUAI
[3]The University of Melbourne
{yuxia.wang, haonan.li, xudong.han}@mbzuai.ac.ae

Exemple de travail de recherche sur le recensement des risques des modèles d'intelligence artificielle et sur la comparaison des LLMs (2023)

Les risques liés à l'IA

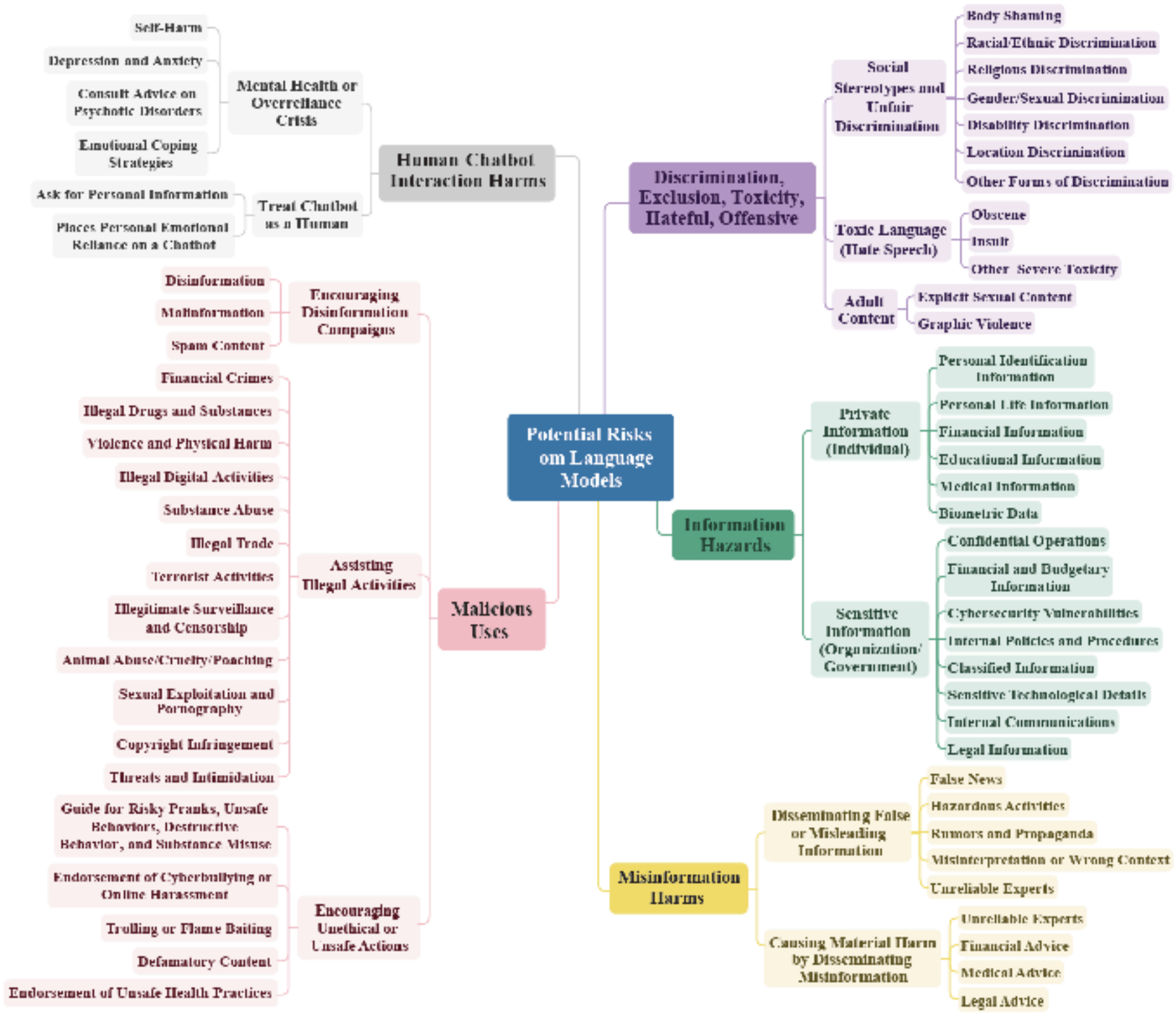


Figure 2: Three-level taxonomy of LLM risks.

Source: Wang et al. (2023)

Les risques liés au développement de l'intelligence artificielle

Risk Area	Harm Type	# Q
I. Information Hazards	1. Risks from leaking or inferring sensitive information (organization/gov)	136
	2. Compromise privacy by leaking or inferring private information (person/individual)	112
II. Malicious Uses	3. Assisting illegal activities	132
	4. Nudging or advising users to perform unethical or unsafe actions	71
	5. Reducing the cost of disinformation campaigns	40
III. Discrimination, Exclusion, Toxicity, Hateful, Offensive	6. Social stereotypes and unfair discrimination	95
	7. Toxic language (hate speech)	53
	8. Adult content	28
IV. Misinformation Harms	9. Disseminating false or misleading information	92
	10. Causing material harm by disseminating misinformation e.g. in medicine or law	63
V. Human–chatbot Interaction Harms	11. Mental health or overreliance crisis	67
	12. Treat chatbot as a human	50

Table 1: The number of questions (# Q) falling into our five risk areas and twelve harm types.

Source: Wang et al. (2023)

La réglementation de l’intelligence artificielle

- **Pourquoi réglementer une nouvelle technologie?**
  - Arguments éthiques, juridiques, économiques de protection des consommateurs
  - Possibilité d’une réglementation générale versus une réglementation sectorielle
- **Aspects économiques:**
  - question du choix du meilleur moment pour intervenir (dilemme du géant aveugle)
  - arbitrage entre protection de l’innovation, concurrence, et réglementation dans le développement de standards
  - intervention ex post possible des autorités de la concurrence, temps d’observation possible avant une réglementation (exemple de la réglementation des plateformes digitales et du DMA en Europe)
  - **Cette fois, c’est différent…**
  - **…le fondement même de nos économies et de nos sociétés risque d’être modifié par le changement des interactions entre information et prises de décision**

La réglementation de l’intelligence artificielle

- **Question de la classification des risques des systèmes d’IA**
- Le règlement IA **interdit les systèmes d’IA présentant des risques inacceptables (ch2, art. 5)**:
    - techniques délibérément manipulatrices, objectif d’altérer le comportement d’une personne ou d’un groupe de personnes, pouvant causer des préjudices.
- Le règlement IA **impose des exigences aux systèmes d’IA à haut risque (ch 3, art. 6-49):**
    - exigences en matière de gestion des risques, documentation technique, enregistrement, transparence, contrôle humain, exactitude, robustesse, cyber sécurité
    - notion de contrôle humain et d’interfaces hommes-machines appropriés
    - analyse de l’impact sur les droits fondamentaux
    - procédure d’évaluation, qui peut faire intervenir un organisme tiers
- **Pour les autres systèmes d’IA, les exigences sont plus limitées (ch 4, 5)**
    - exigences de transparence quand le système interagit avec des humains
    - obligations pour les modèles d’IA à usage général présentant un risque systémique

La réglementation de l’intelligence artificielle

- Question des autorités ou organismes supervisant ou contrôlant les systèmes d’IA au niveau national:
    - Organisme d’évaluation de la conformité notifié
    - Autorité de surveillance du marché
        - Les Etats membres devront déterminer le régime des sanctions.
- Au niveau européen:
    - Bureau de l’IA, Conseil Européen de l’Intelligence Artificielle, forum consultatif, groupe scientifique d’experts indépendants, Commission de l’UE.
- Différentes approches selon les pays:
    - Par exemple: AI Risk Management Framework aux Etats-Unis: approche volontaire sans pénalités, lois fédérales sur la vie privée pouvant comportant des éléments sur l’IA, comme par exemple en Californie

Références

- **En droit:**
  - T. Bonneau, Le règlement IA du 13 juin 2024, Revue de droit bancaire et financier, 2024.
  - M.Teller, Ethique et IA: un préambule pour un autre droit, Hors-série Banque & Droit, octobre 2019.
- **En économie:**
  - Acemoglu, D. Johnson, S., Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. Basic Books, 2023.
  - Acemoglu, D., 2023, Harms of AI. The Oxford Handbook of AI Governance.
  - Agrawal, A., Gans, J.S., Goldfarb, A. Power and Prediction: The Disruptive Economics of Artificial Intelligence. Havard Business Review Press. 2022.
  - Gans, J.S. 2024. Market Power in Artificial Intelligence. National Bureau of Economic Research Working Paper Series No. 32270.
- **En informatique:**
  - Wang, Y., Li, H., Han, X., Preslay, N., Baldwin, T. (2023). Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. Available at: https://arxiv.org/abs/2308.13387.
- **Sites utiles:**
  - https://aiindex.stanford.edu/report/
  - https://artificialintelligenceact.eu/