

Intelligence Artificielle et Robotique des fondements scientifiques aux questions éthiques

Raja Chatila

Institut des Systèmes Intelligents et de Robotique (ISIR)

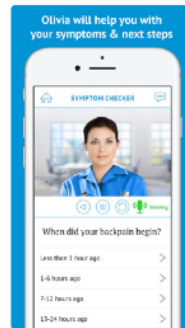
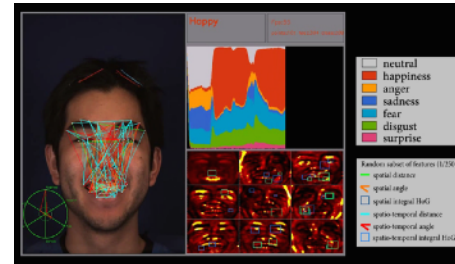
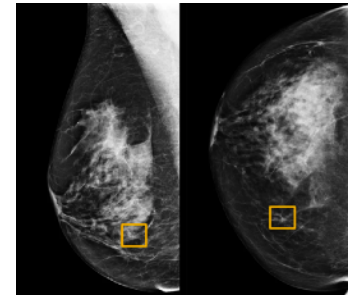
Sorbonne Université, faculté des Sciences et d'Ingénierie, Paris, France

Raja.Chatila@sorbonne-universite.fr

- Qu'est que l'IA? Différentes approches
- Systèmes symboliques
- Systèmes connexionnistes - Réseaux de neurones
- IA générative
- Limitations
- Ethique
- Enjeux éthiques et sociétaux de l'IA
- Conclusion

Nombreuses Applications de l'IA et de la Robotique

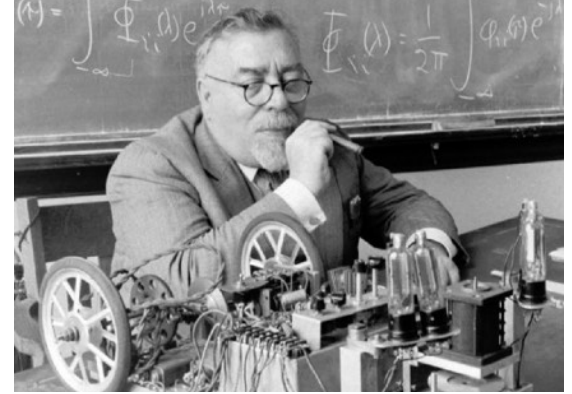
- Transports logistique
- Santé
- Fabrication manufacturière
- Gestion de déchets
- Réseaux électriques
- Agriculture
- Services personnels et assistance
- e-Commerce, publicité, recommandations
- Recrutement & management
- Assurance & finance
- Justice
- Sécurité
- Armement
- ...



Intérêt de l'IA et de la Robotique

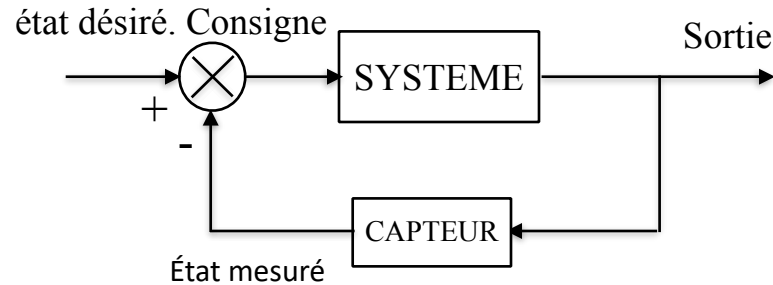
- Augmenter la productivité grâce à l'automatisation de processus matériels ou logiciels
- Réaliser des tâches répétitives ou non réalisables (trop coûteuses, trop difficiles, trop dangereuses, voire impossibles pour les humains)
- Exploiter les données massives (images, textes, sons, données scientifiques, ...)
- Créer de nouveaux services grâce au traitement des données et à la connectivité

La Cybernétique



Norbert Wiener (1894 - 1964)

- “Cybernetics or Control and Communication in the Animal and the Machine “(1948).
- Homéostat, boucle fermée, systèmes automatiques.



Les machines peuvent-elles penser ?

1936: Machine Universelle de Turing -
modèle formel de l'ordinateur

1950 : "Computing Machinery and Intelligence"



Alan Turing (1912-1954)

Les machines peuvent-elles penser ?

1936: Machine Universelle de Turing -
modèle formel de l'ordinateur

1950 : "Computing Machinery and Intelligence"



Alan Turing (1912-1954)

VOL. LIX. No. 236.]

[October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?
Now suppose X is actually A, then A must answer. It is A's

28

433

Les machines peuvent-elles penser ?

1936: Machine Universelle de Turing -
modèle formel de l'ordinateur

1950 : "Computing Machinery and Intelligence"



Alan Turing (1912-1954)

VOL. LIX. No. 236.]

[October, 1950

MIND
A QUARTERLY REVIEW
OF

I propose to consider the question "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words but this attitude is dangerous ...

...

The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people a man (A) a woman (B) and an interrogator (C) who may be of either sex

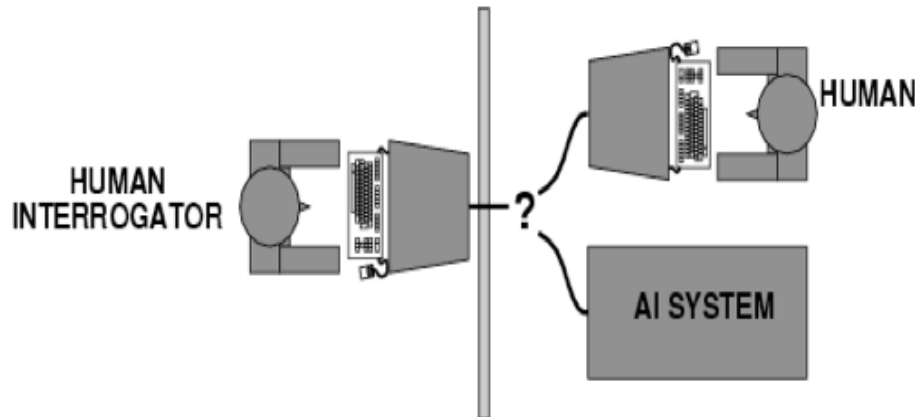
....

Les machines peuvent-elles penser ?

1936: Machine Universelle de Turing -
modèle formel de l'ordinateur

1950 : "Computing Machinery and Intelligence"

le jeu de l'imitation (Test de Turing)



Alan Turing (1912-1954)

VOL. LIX. No. 236.]

[October, 1950

MIND
A QUARTERLY REVIEW
OF

I propose to consider the question "**Can machines think?**" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words but this attitude is dangerous ...

...

The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people a man (A) a woman (B) and an interrogator (C) who may be of either sex

....

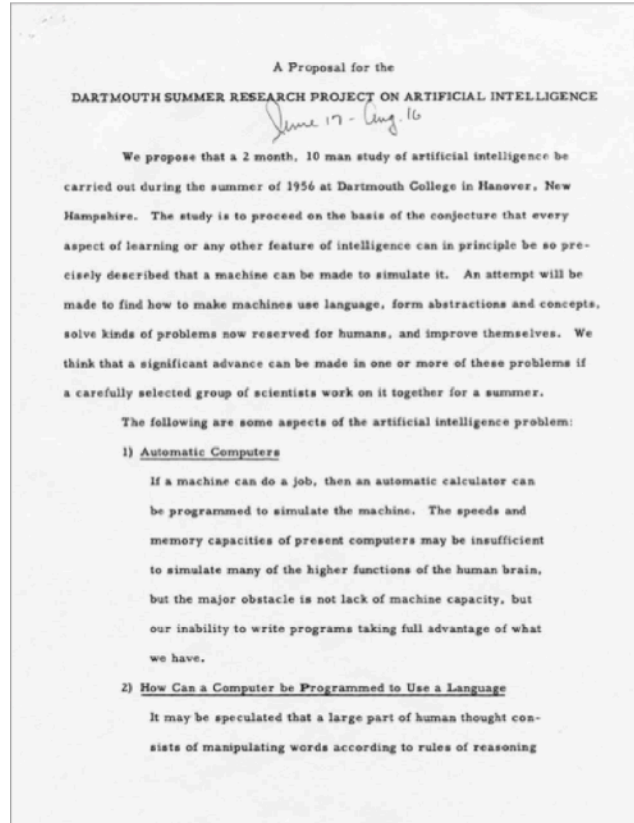
Naissance de l'Intelligence Artificielle: La conférence de Dartmouth College, 1956



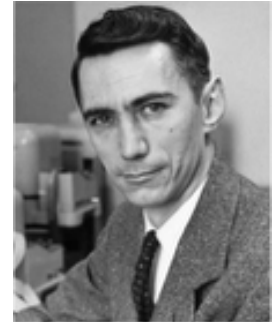
John McCarthy



Marvin Minsky



Nathaniel Rochester



Claude Shannon

Naissance de l'Intelligence Artificielle: La conférence de Dartmouth College, 1956



John McCarthy

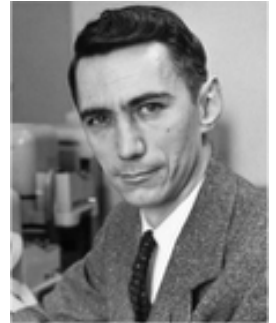


Marvin Minsky

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language form abstractions and concepts solve kinds of problems now reserved for humans and improve themselves.



Nathaniel Rochester

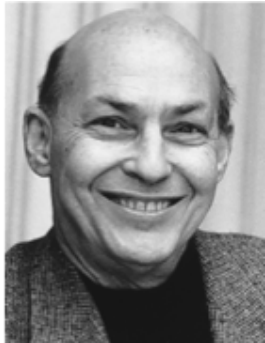


Claude Shannon

Naissance de l'Intelligence Artificielle: La conférence de Dartmouth College, 1956



John McCarthy



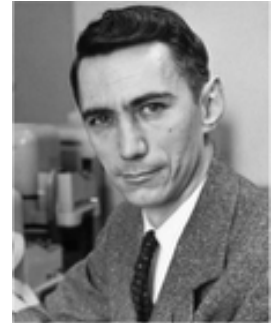
Marvin Minsky

The study is to proceed on the basis of the conjecture that every aspect of learning or any other intelligence can in principle be precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language form abstractions and concepts solve kinds of problems now reserved for humans and improve themselves.

ALGORITHMES



Nathaniel Rochester



Claude Shannon

Grands domaines de l'IA

Apprentissage Machine

Systèmes connexionnistes

Apprentissage

Profond

Supervisé / non-supervisé

Apprentissage
par renforcement

Robotique

Perception

Décision

Action

Interaction

IA Symbolique

Représentation des connaissances

Raisonnement logique ou probabiliste

Résolution de problèmes, Planification, ...

Automatique
Mécanique
Electronique
Informatique
Temps réel
Sûreté
Systèmes

Le robot

Machine matérielle située et agissant dans le monde réel

L'intelligence vue comme l'interaction rationnelle et efficiente avec le monde

- **Perception**
- **Action et mouvement**
- **décision et réactivité**
- **Communication et interaction**
- **apprentissage**



Capacités intégrées développées à divers degrés de complexité permettant des niveaux d'automatisation [*autonomie*] différents



Boston Dynamics



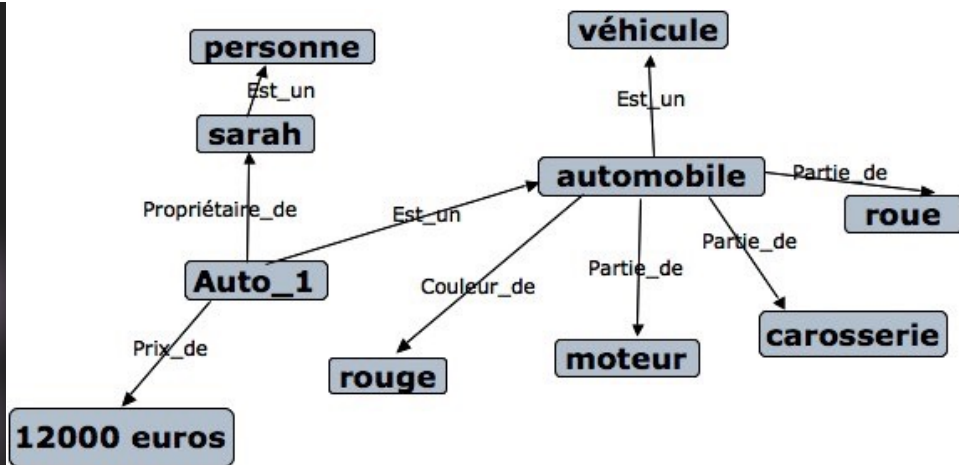
Boston Dynamics



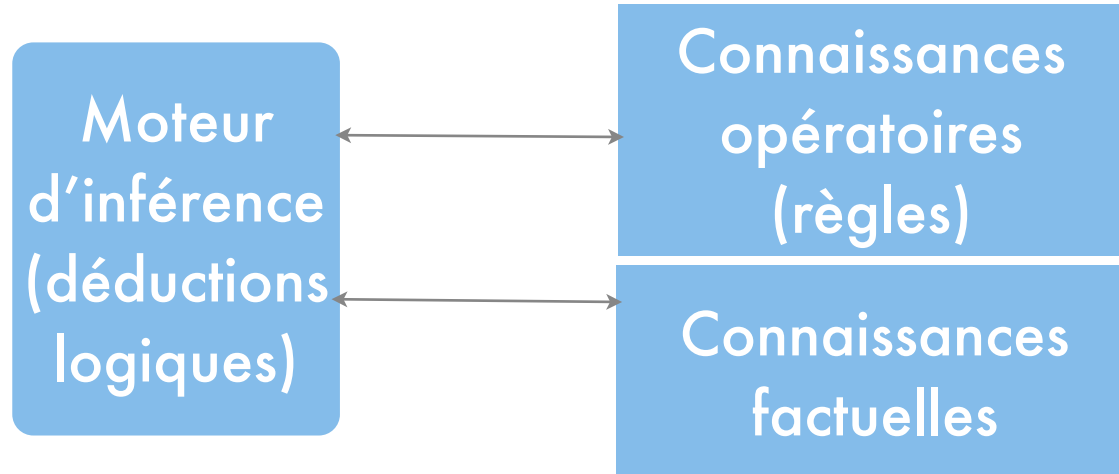
Boston Dynamics

Exemples de méthodes de l'IA Symbolique

Ontologies, Réseau de connaissances

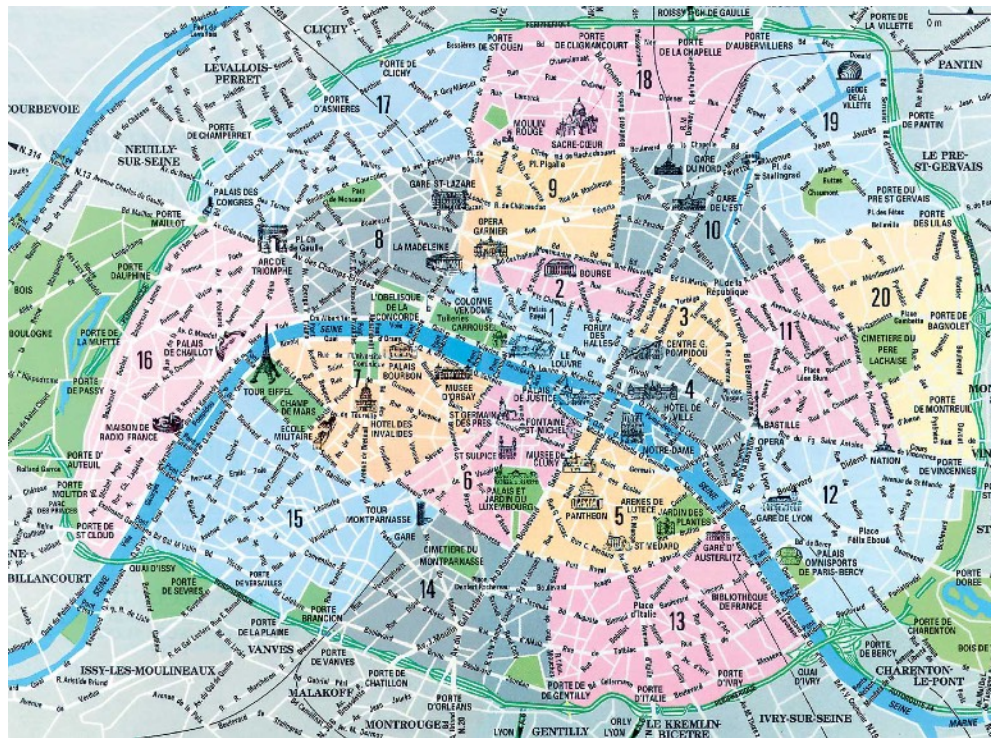
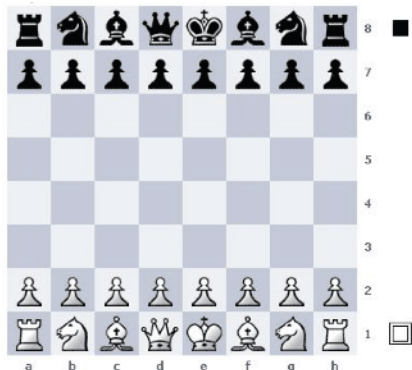


Systemes à base de règles

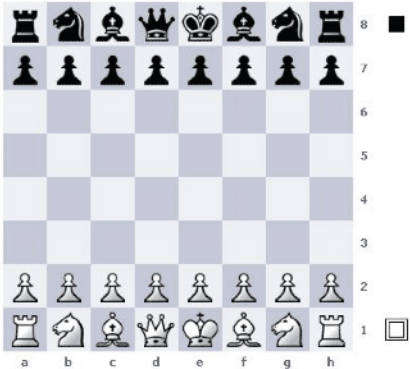


Applications: systèmes experts

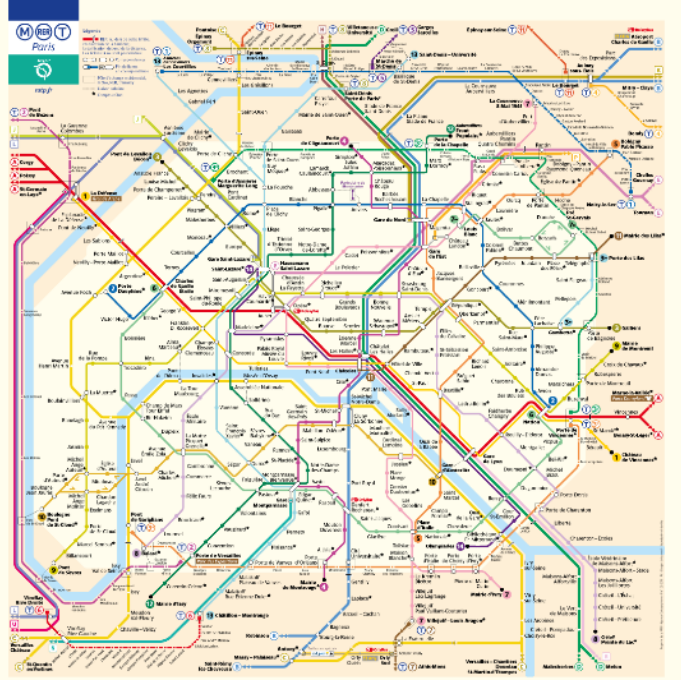
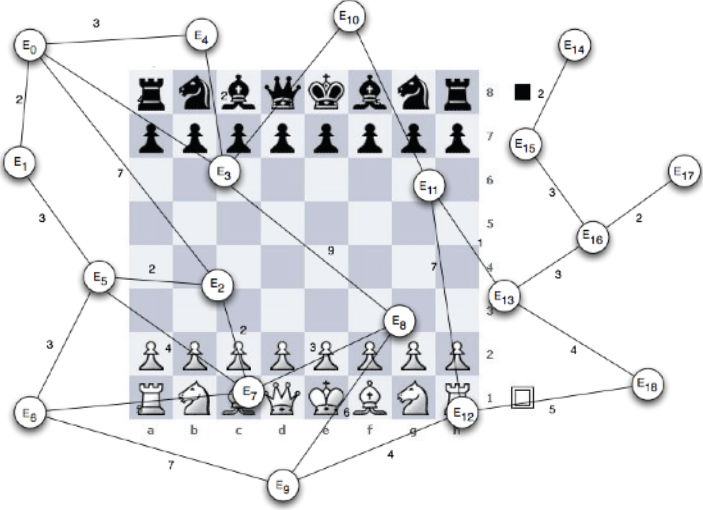
Planification d'actions



Planification d'actions

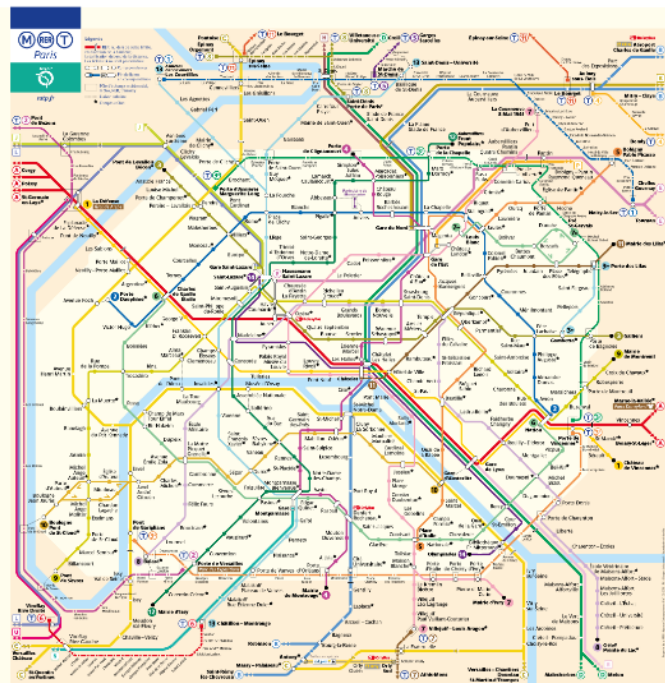
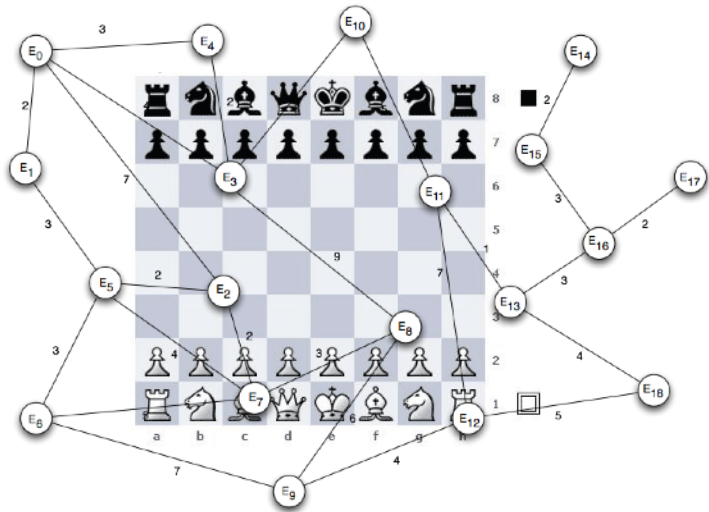


Planification d'actions



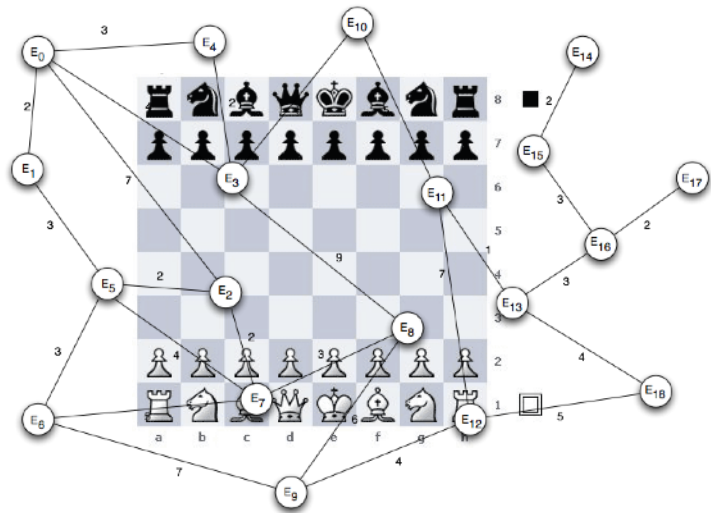
Planification d'actions

Etat: présence dans une station de métro donnée

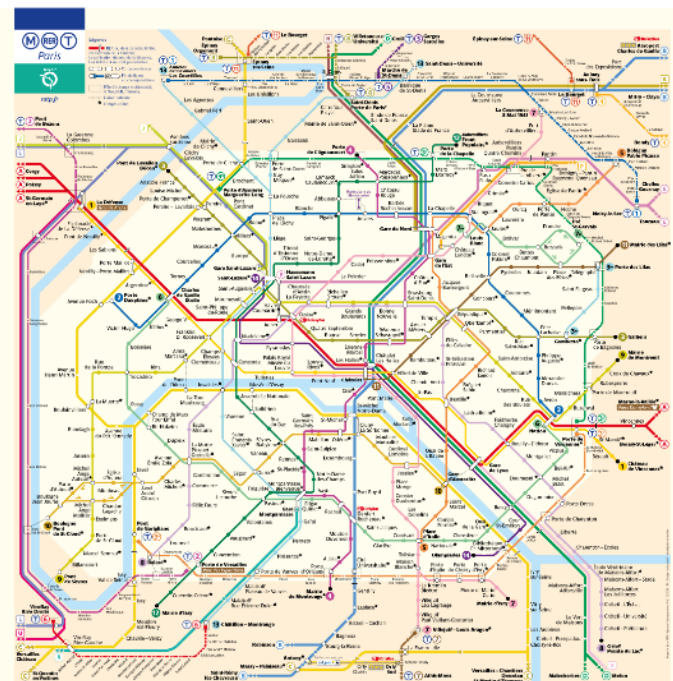


Planification d'actions

Etat: position de toutes les pièces sur l'échiquier

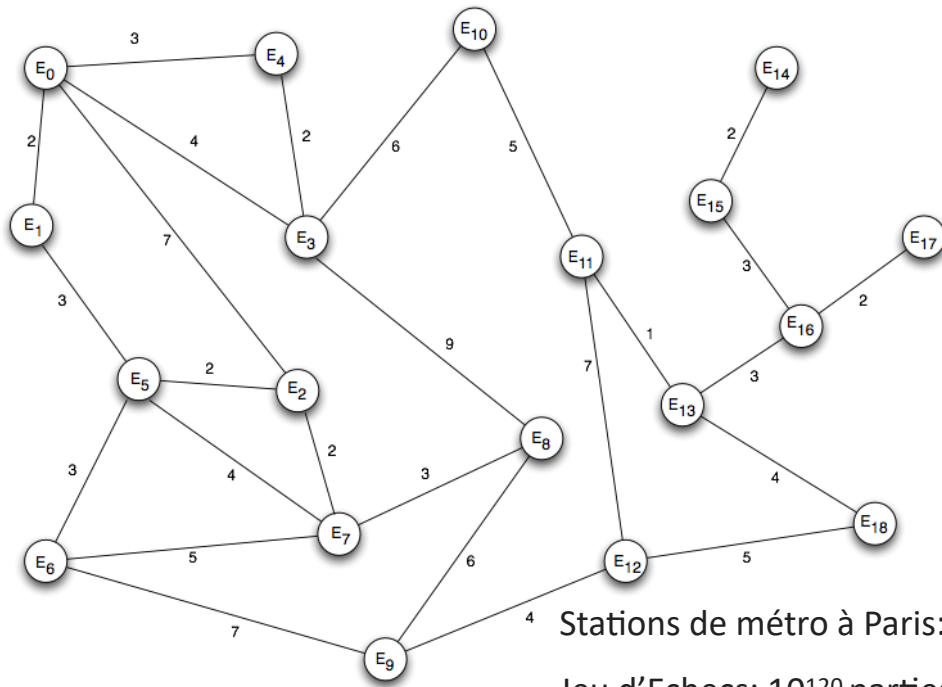


Etat: présence dans une station de métro donnée

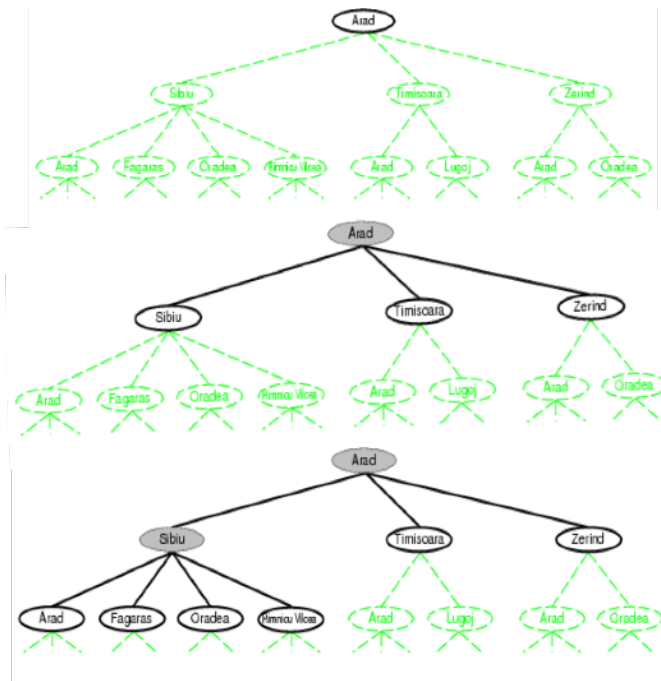


Décision et résolution de problèmes: comment aller de l'état E_0 à l'état E_n ? **Gérer la complexité**

Graphe d'états



Arbre de recherche



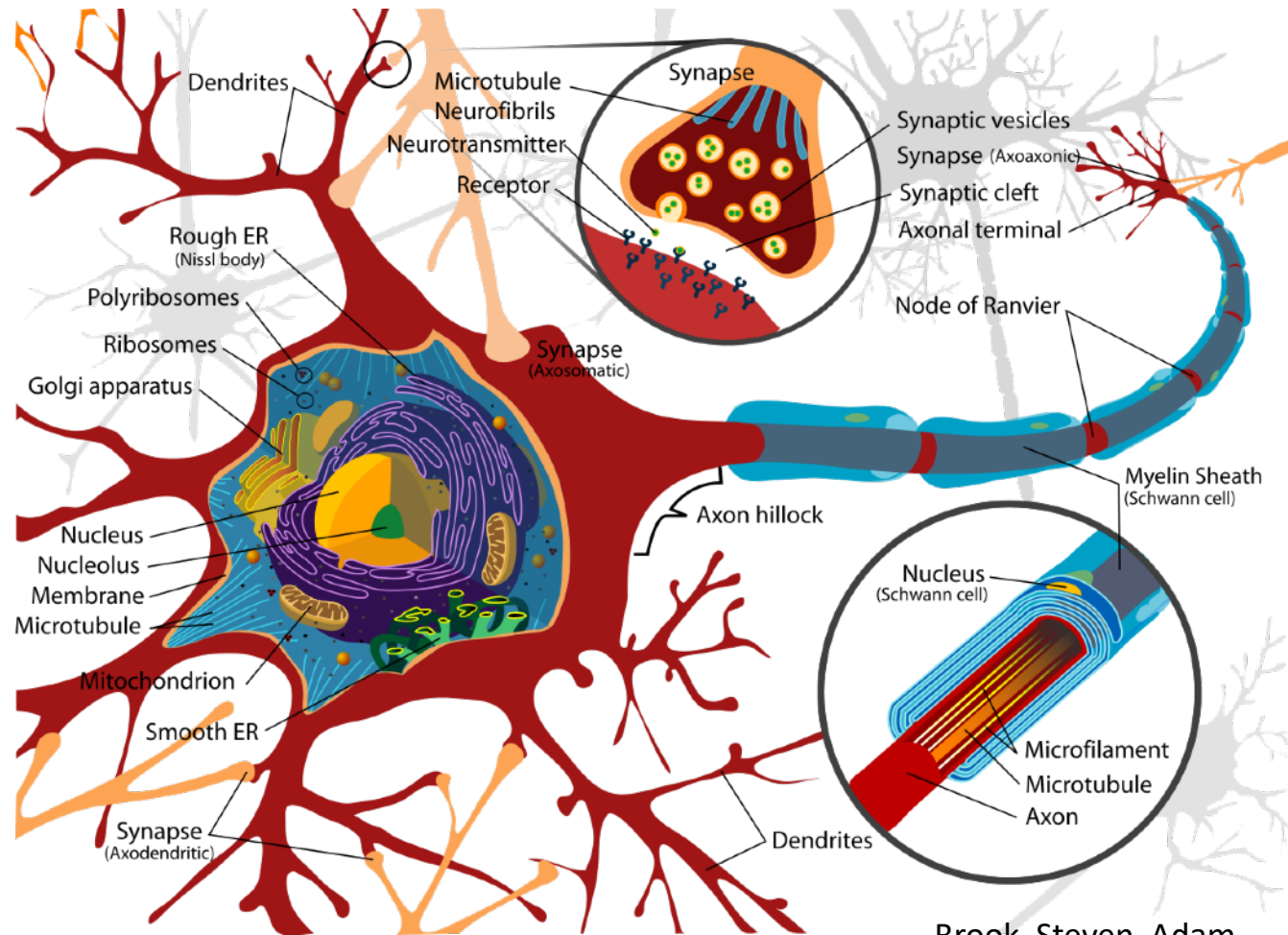
Stations de métro à Paris: 309

Jeu d'Echecs: 10^{120} parties différentes possibles (dont 10^{50} légales)

Jeu de Go: 10^{600} parties possibles

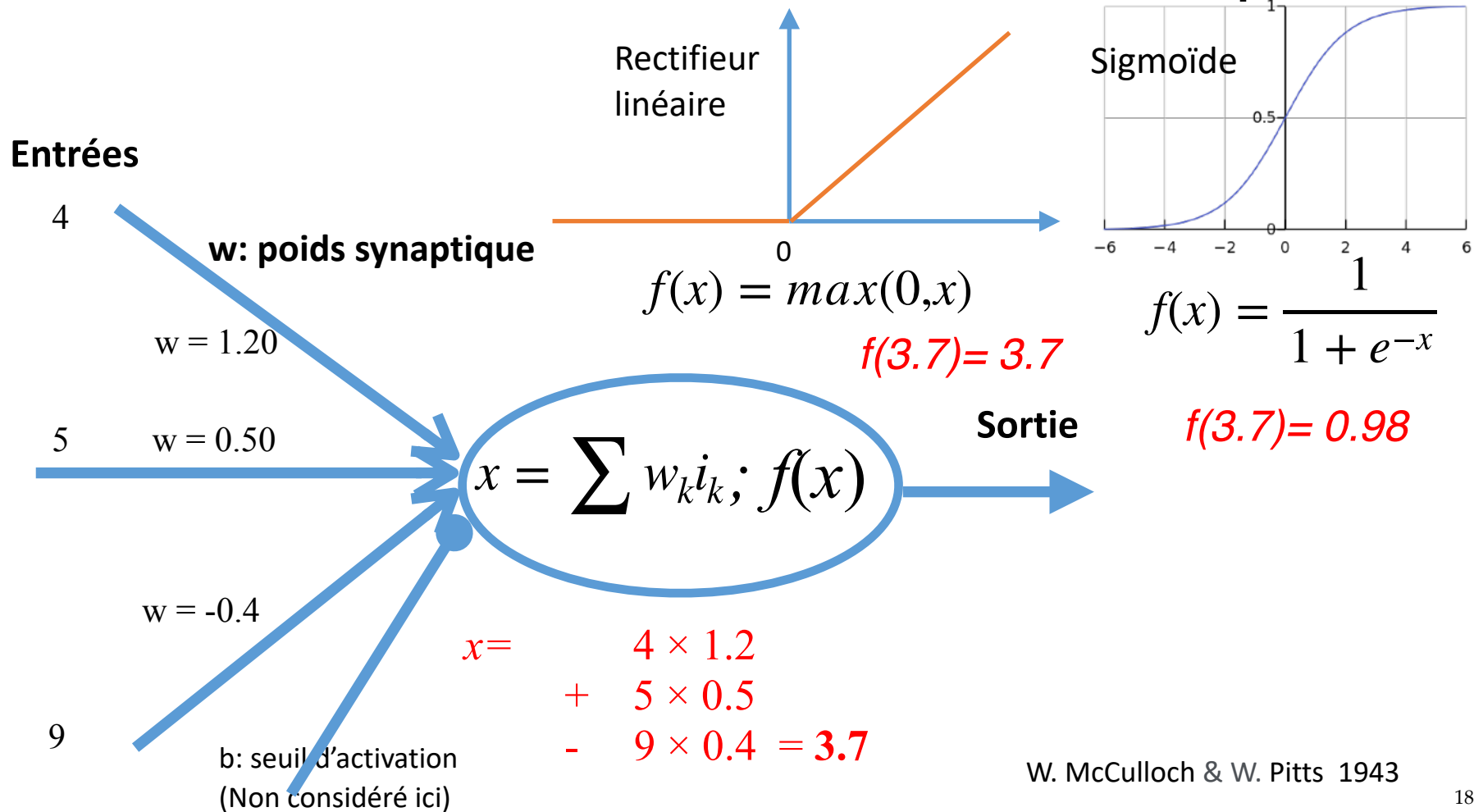
Approches connexionnistes

Le Neurone Naturel



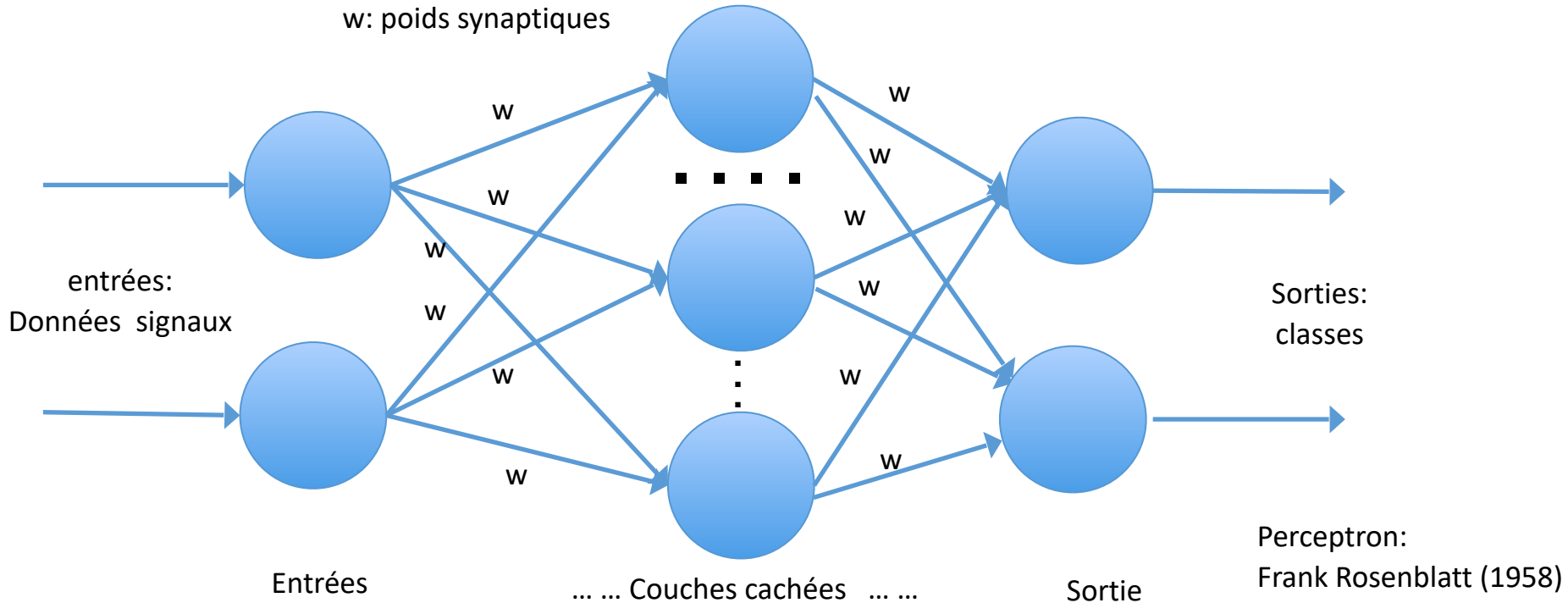
Brook Steven Adam

Le neurone formel : une fonction mathématique



Réseau de Neurones

Millions, milliards, centaines de milliards de paramètres selon la taille du réseau



Apprentissage Machine

Traitement statistique de données et classification:

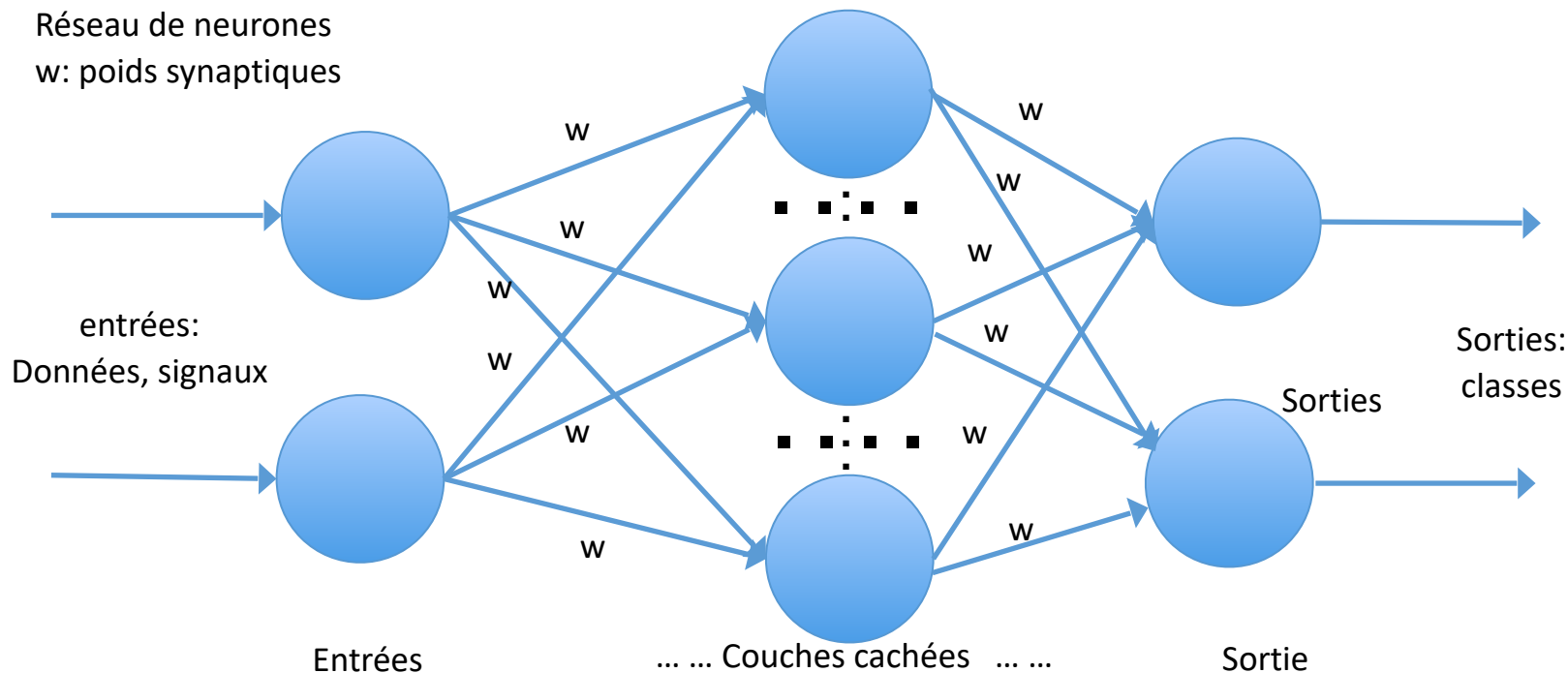
- Distributions de probabilité et corrélations
- Réseaux de neurones formels comme classifieurs
- Algorithmes d'optimisation

- **Apprentissage supervisé** : la réponse correcte est fournie par une connaissance *a priori*

- **Apprentissage non-supervisé** : recherche de régularités dans les données

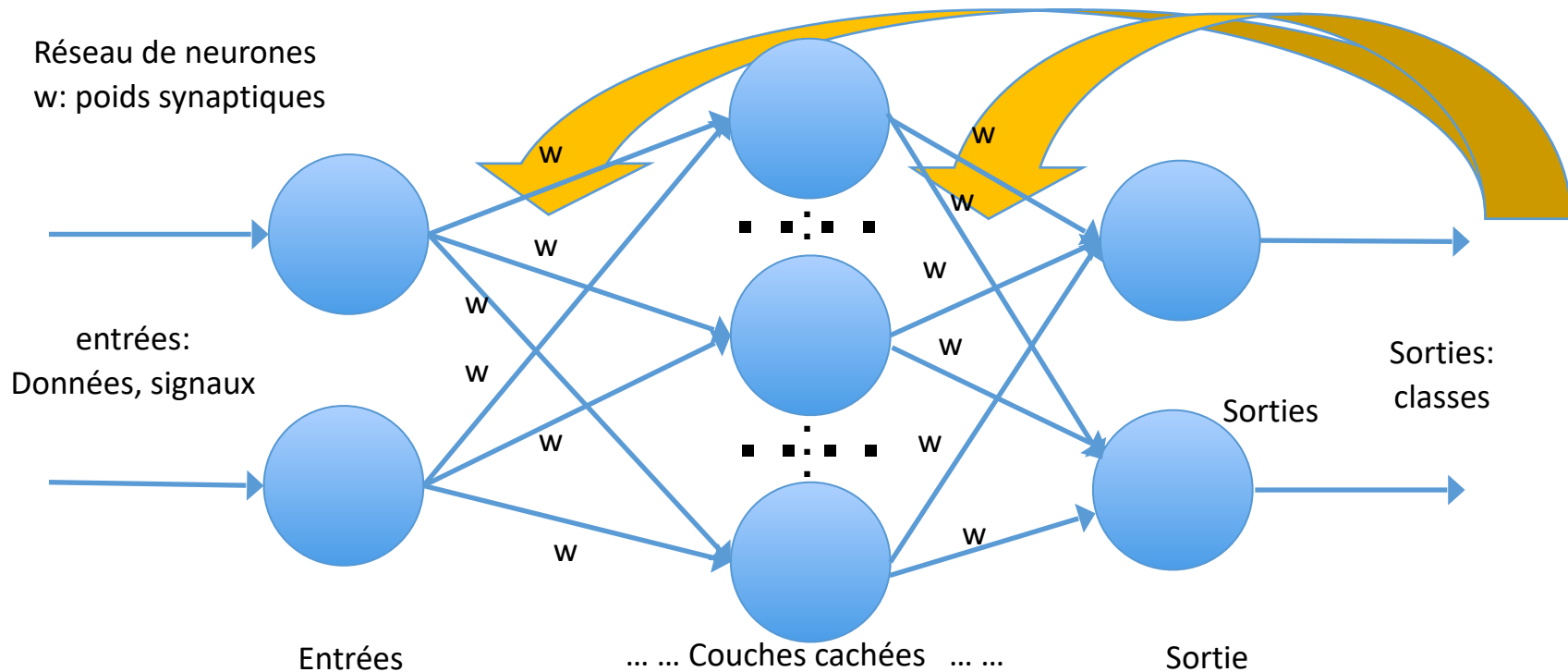
- **Apprentissage par renforcement** : recherche des actions les plus prometteuses en optimisant une fonction objectif (récompense cumulée)

Apprentissage (ou entraînement) supervisé



Millions, milliards, centaines de milliards de paramètres selon la taille du réseau

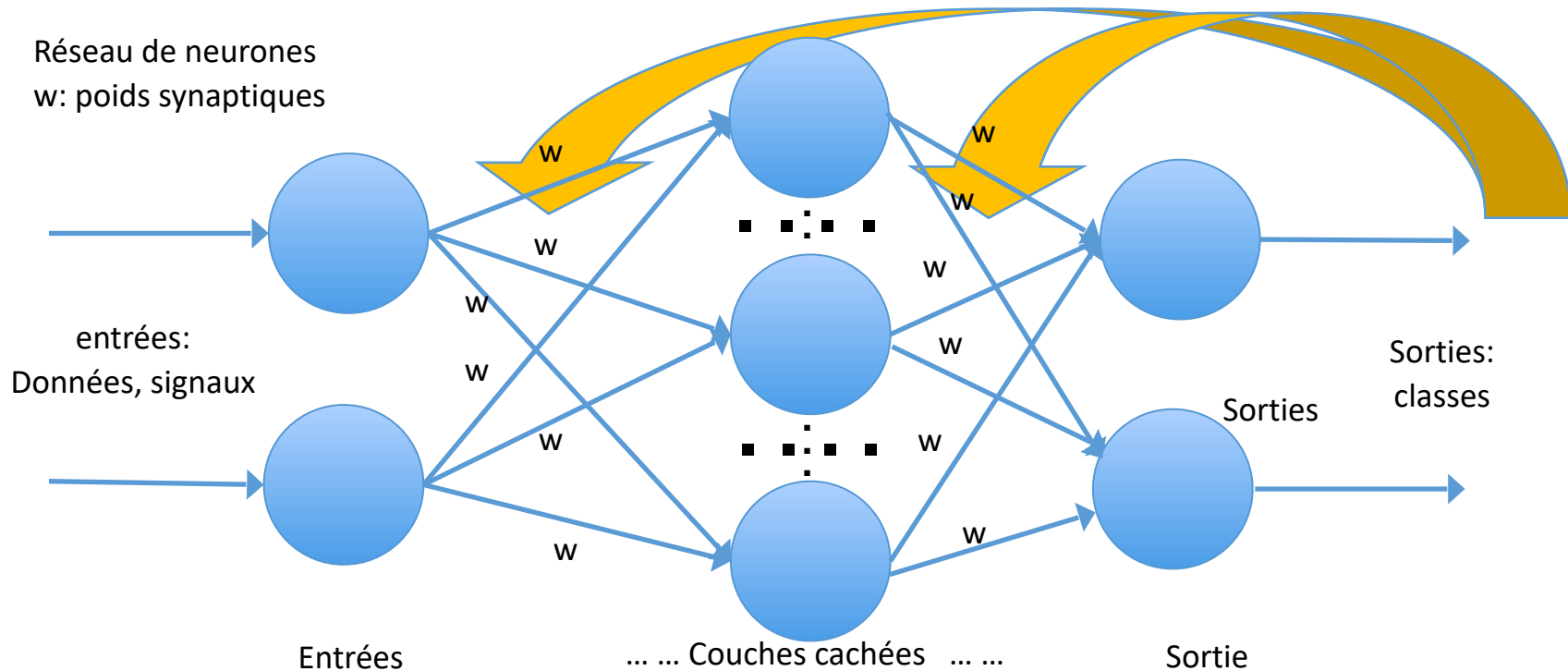
Apprentissage (ou entraînement) supervisé



Millions, milliards, centaines de milliards de paramètres selon la taille du réseau

Apprentissage (ou entraînement) supervisé

Optimisation (p.ex. Descente de gradient)
Propagation des erreurs (loss function)
Ajustement itératif des poids synaptiques w_i pour obtenir les sorties désirées



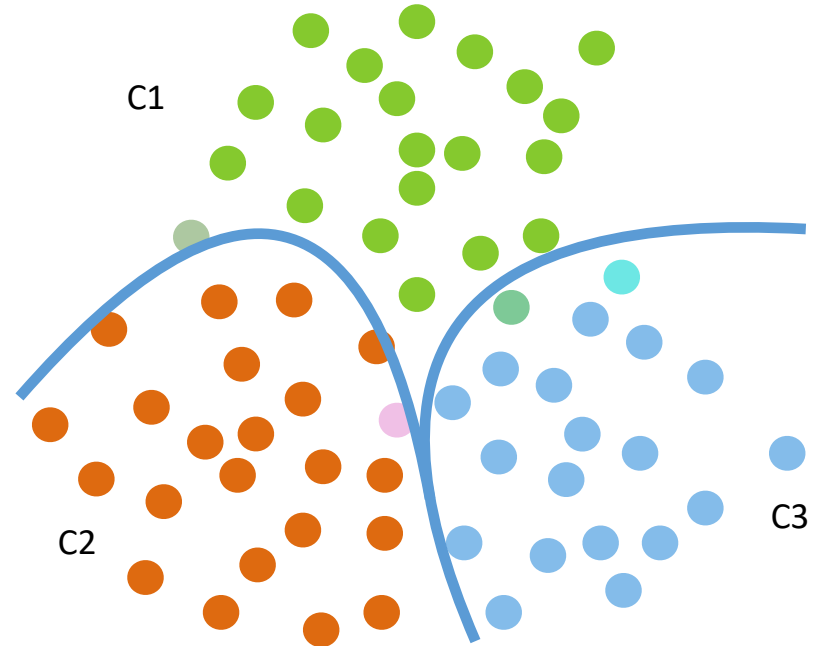
Millions, milliards, centaines de milliards de paramètres selon la taille du réseau

Apprentissage non supervisé

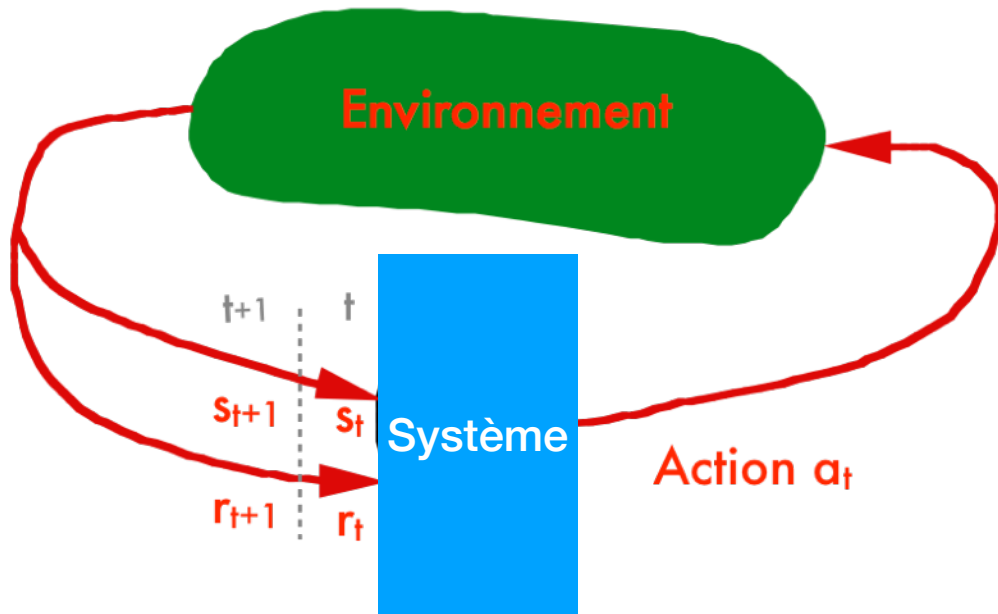
Recherche de régularités/similarités
dans les données.

Méthodes:

Regroupement, classification



Apprentissage par renforcement



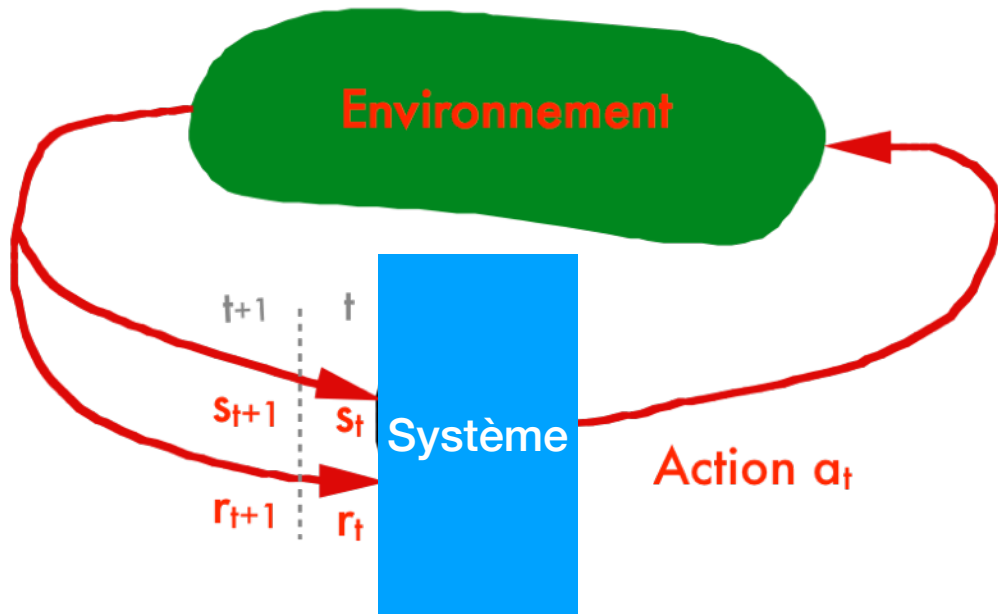
**Path Integral
Guided Policy Search**

Yevgen Chebotar, Mrinal Kalakrishnan, Ali Yahya
Adrian Li, Stefan Schaal, Sergey Levine

Maximiser la somme des récompenses obtenues

$$Q(s, a) = R(s) + \gamma \sum_{s'} P(s, a, s') \max_{a'} Q(s', a')$$

Apprentissage par renforcement



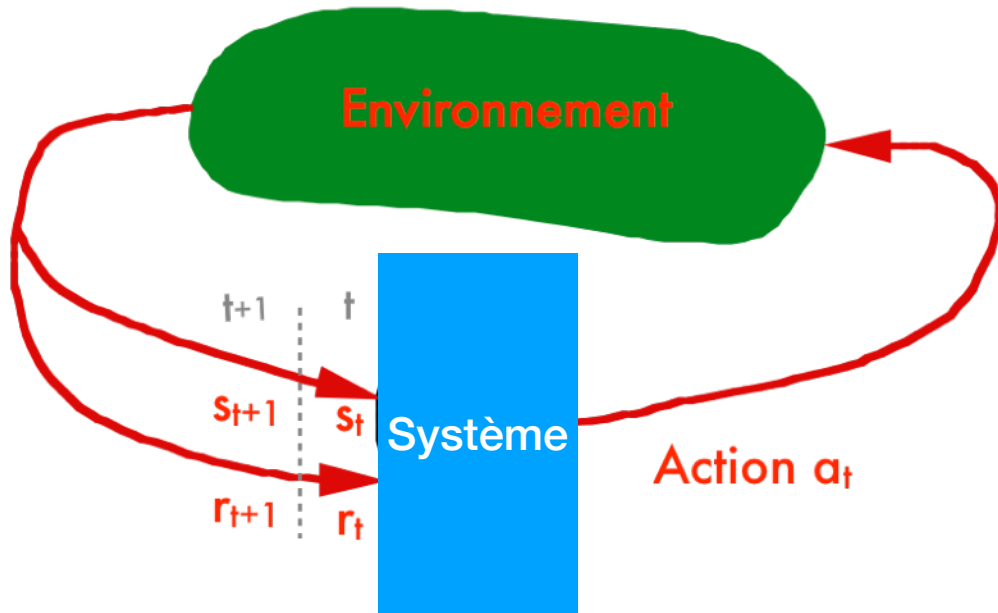
**Path Integral
Guided Policy Search**

Yevgen Chebotar, Mrinal Kalakrishnan, Ali Yahya
Adrian Li, Stefan Schaal, Sergey Levine

Maximiser la somme des récompenses obtenues

$$Q(s, a) = R(s) + \gamma \sum_{s'} P(s, a, s') \max_{a'} Q(s', a')$$

Apprentissage par renforcement



**Path Integral
Guided Policy Search**

Yevgen Chebotar, Mrinal Kalakrishnan, Ali Yahya
Adrian Li, Stefan Schaal, Sergey Levine

Maximiser la somme des récompenses obtenues

$$Q(s, a) = R(s) + \gamma \sum_{s'} P(s, a, s') \max_{a'} Q(s', a')$$

Systemes d'IA Prédictifs, Systemes d'IA Génératifs

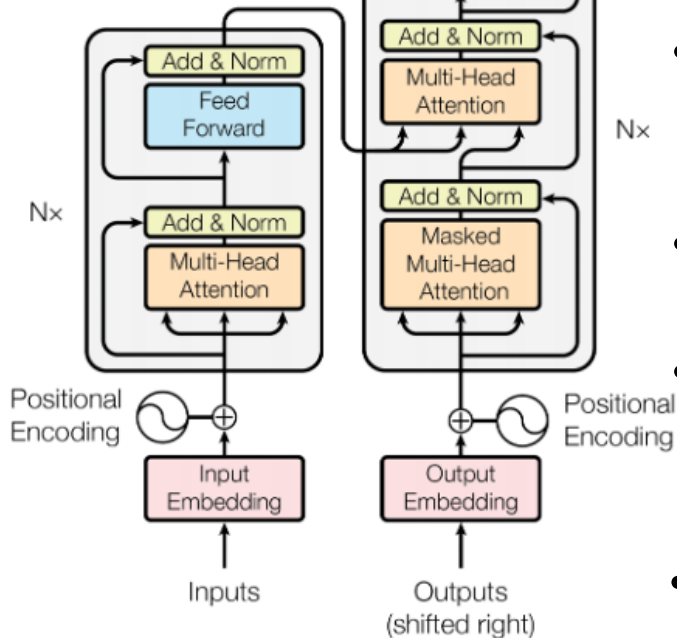
Systemes Prédictifs: calculent l'appartenance de l'entrée à une classe (corrélacion statistique)

Systemes Génératifs: produisent une sortie qui se corrèle bien avec les données d'entrée

Transformers - Modèles de Fondation

Modèles à usage général, IA générative

Vaswani et al. Attention is all you need
<https://arxiv.org/abs/1706.03762>
(NeurIPS 2017)



Focalisation de l'attention: contexte

Structures de données séquentielles (ex: langue naturelle)

- Apprentissage non-supervisé de probabilités de distribution à partir de textes composés de séquences de symboles de longueurs variables
- Utilisation de “tokens” pour coder des séquences de caractères
- Generative Pretrained Transformers
 - GPT-3: 4 096 tokens ; 175 Mds de paramètres;
 - GPT-4 : 32 768 tokens (~ 50 pages)
- ChatGPT: Interface utilisateur - Génération interactive de texte à partir d'invites (prompts)

IA Générative

Grands Modèles de Langage

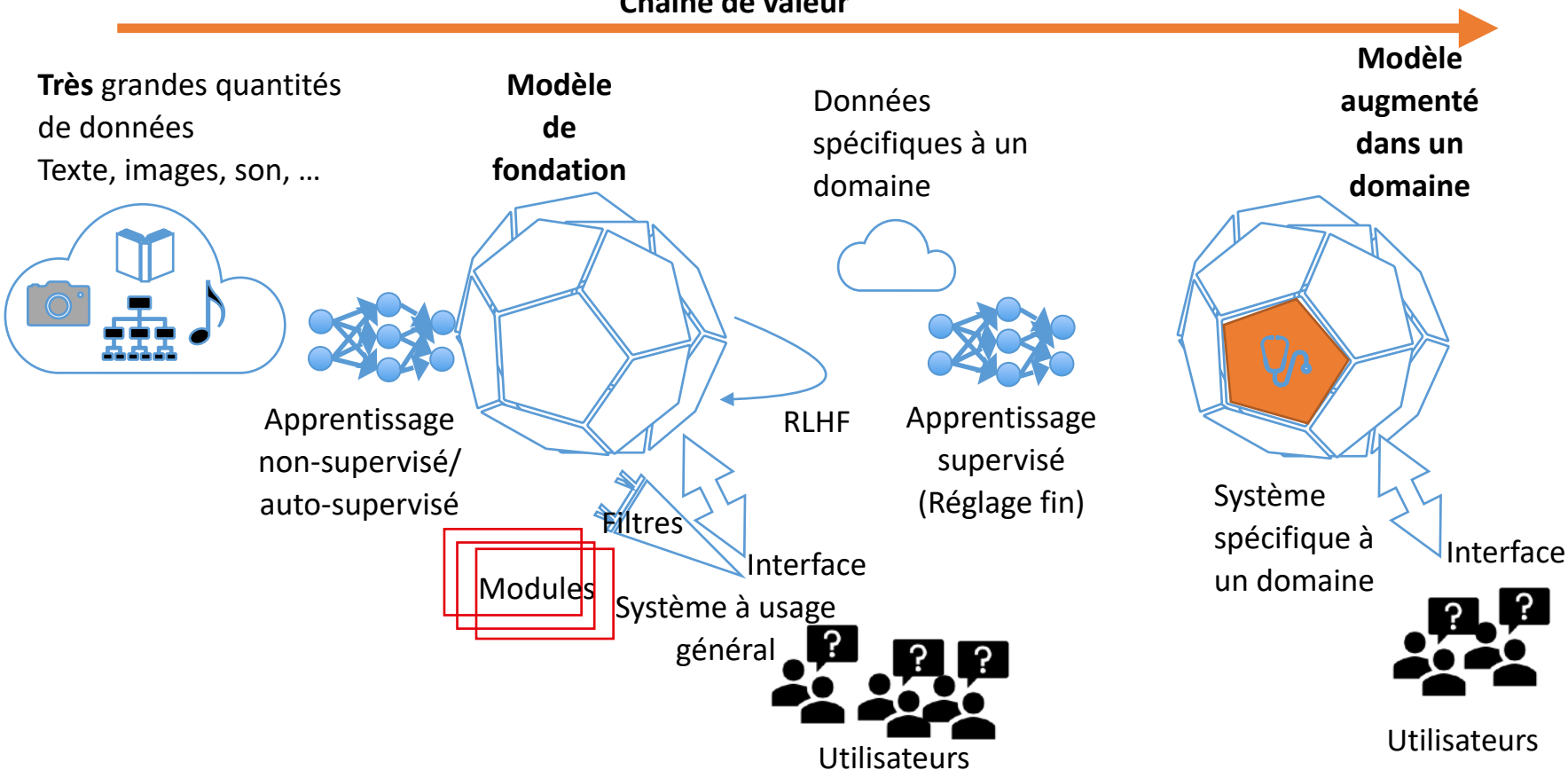
- Utilisation de l'architecture Transformers et apprentissage sur de grandes quantités de données textuelles présentes sur internet
- GPT3 (OpenAI): 175 Mds de paramètres; LAMDA (Google) 137Mds ; BLOOM (Hugging Face) 176 Mds (59 langues/langages) Llama (Meta), ...
- Ré-apprentissage sur des données dans des domaines d'application spécifiques
- ChatGPT: Interface utilisateur - Génération interactive de texte à partir d'invites (prompts)

Language Models are Unsupervised Multitask Learners. Alec Radford and Jeffrey Wu and R. Child and David Luan and Dario Amodei and Ilya Sutskever 2019.

Language Models are Few-Shot Learners. Brown Tom et al. (31 authors). Neurips 2020

Modèles et Systèmes génératifs

Chaîne de valeur



Limites et questions éthiques

Limites de l'apprentissage statistique: Biais des données

Wrong



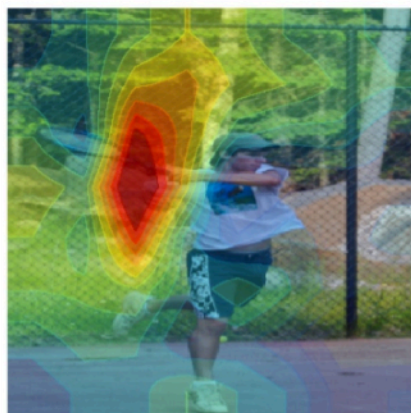
Baseline:
*A **man** sitting at a desk with
a laptop computer.*

Right for the Right
Reasons



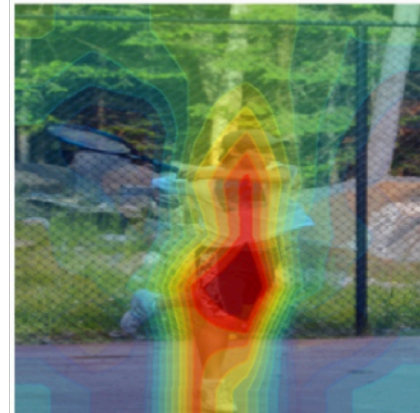
Our Model:
*A **woman** sitting in front of a
laptop computer.*

Right for the Wrong
Reasons



Baseline:
*A **man** holding a tennis
racquet on a tennis court.*

Right for the Right
Reasons

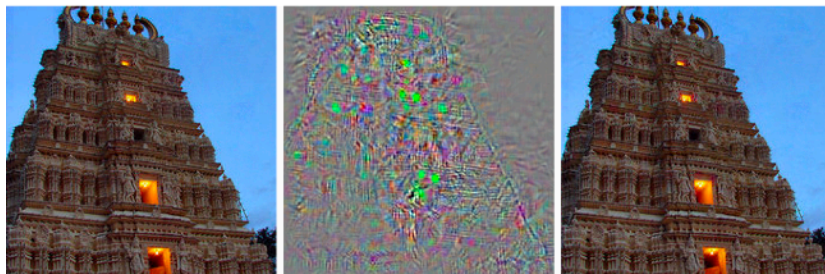
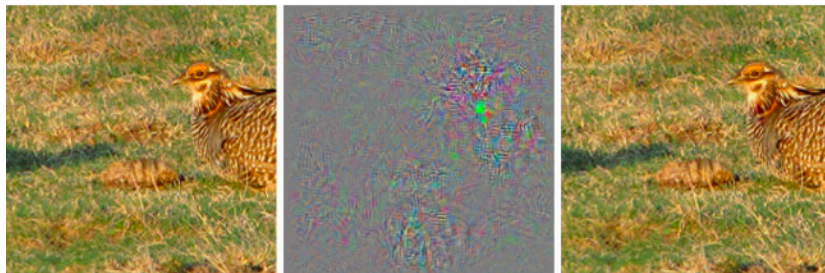
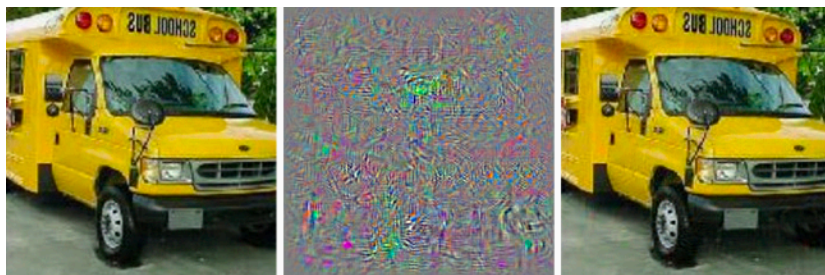


Our Model:
*A **man** holding a tennis
racquet on a tennis court.*

Women also Snowboard: Overcoming Bias in Captioning Models.

Lisa Anne Hendricks Kaylee Burns Kate Saenko Trevor Darrell Anna Rohrbach. ECCV 2018

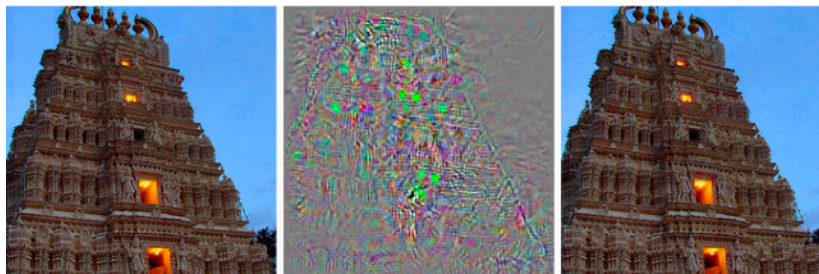
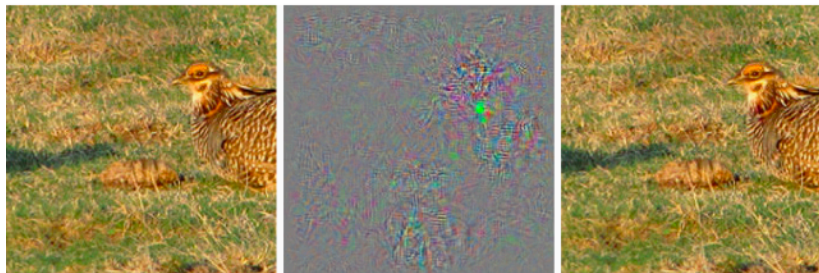
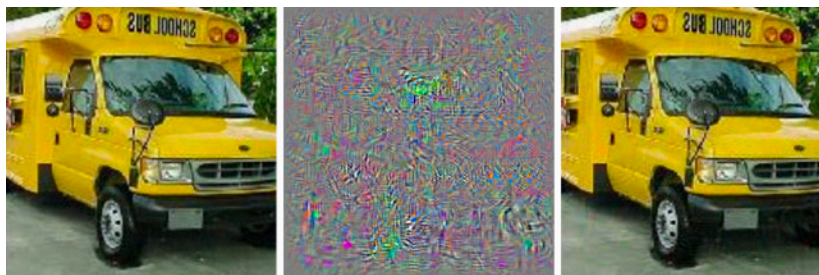
Limites de l'apprentissage statistique: Sensibilité au bruit



Intriguing properties of neural networks
Christian Szegedy Wojciech Zaremba Ilya
Sutskever Joan Bruna Dumitru Erhan Ian
Goodfellow Rob Fergus

<https://doi.org/10.48550/arXiv.1312.6199>

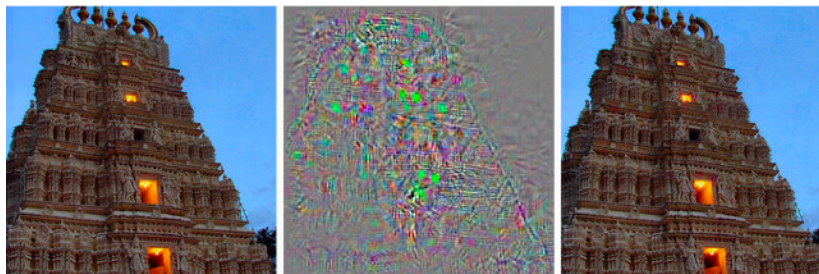
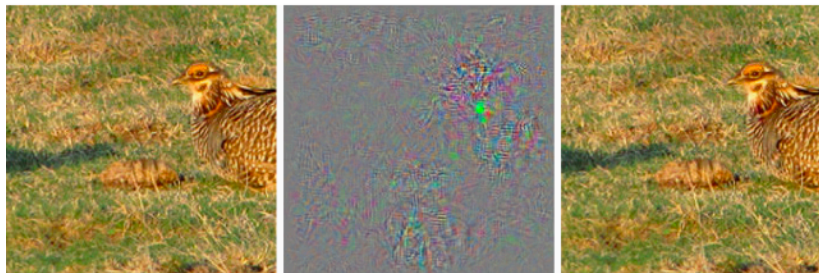
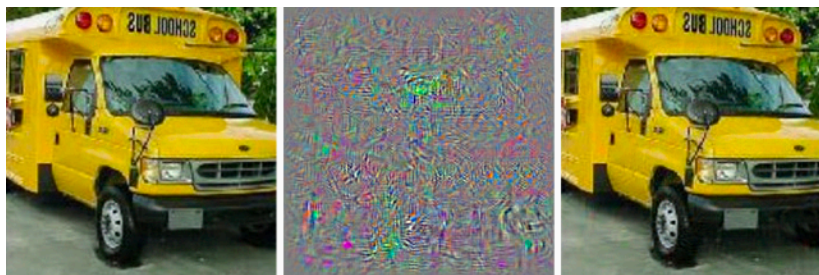
Limites de l'apprentissage statistique: Sensibilité au bruit



Intriguing properties of neural networks
Christian Szegedy Wojciech Zaremba Ilya
Sutskever Joan Bruna Dumitru Erhan Ian
Goodfellow Rob Fergus
<https://doi.org/10.48550/arXiv.1312.6199>

Images initiales.
Interprétations
correctes du système

Limites de l'apprentissage statistique: Sensibilité au bruit



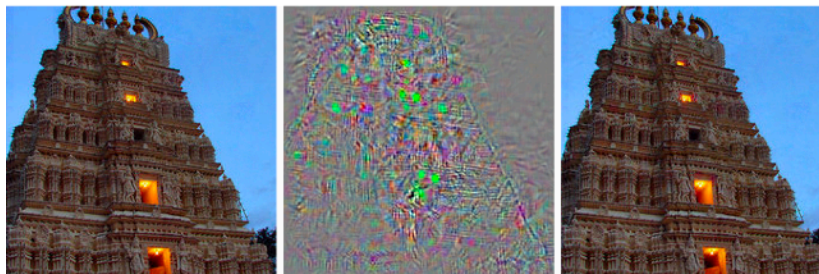
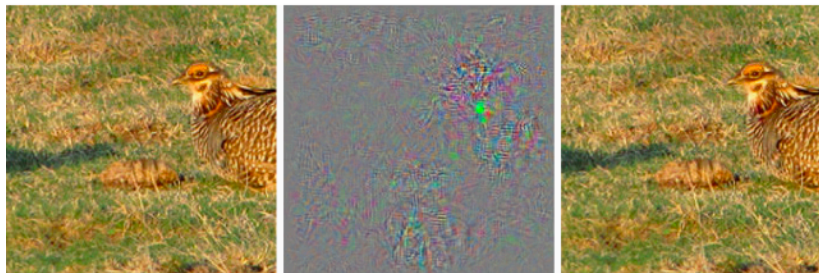
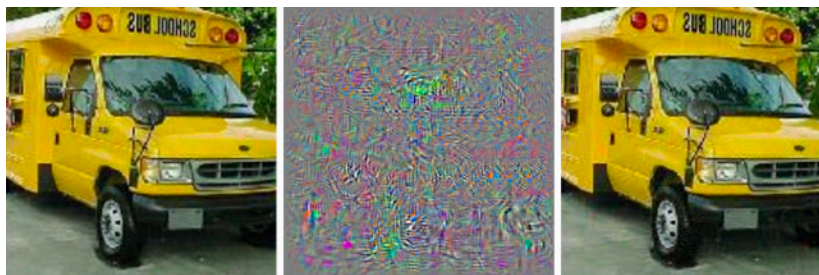
Intriguing properties of neural networks
Christian Szegedy Wojciech Zaremba Ilya
Sutskever Joan Bruna Dumitru Erhan Ian
Goodfellow Rob Fergus

<https://doi.org/10.48550/arXiv.1312.6199>

Images initiales.
Interprétations
correctes du système

+ bruit

Limites de l'apprentissage statistique: Sensibilité au bruit



Intriguing properties of neural networks
Christian Szegedy Wojciech Zaremba Ilya
Sutskever Joan Bruna Dumitru Erhan Ian
Goodfellow Rob Fergus

<https://doi.org/10.48550/arXiv.1312.6199>

Images initiales.
Interprétations
correctes du système

+ bruit

Interprétation
du système:
“autruche”

Limites de l'apprentissage statistique: Robustesse

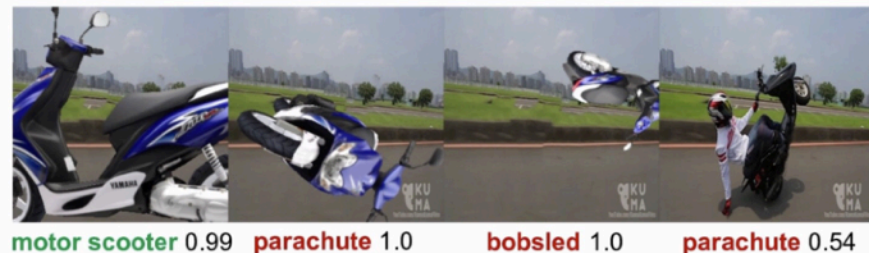
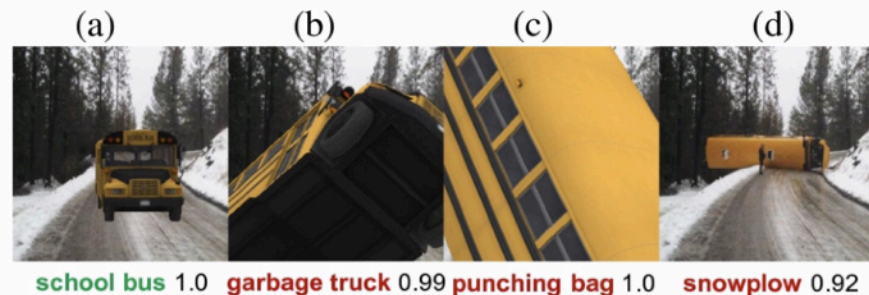


Classé comme



Des perturbations mineures de l'image entraînent des erreurs de classification.

Robust Physical-World Attacks on Deep Learning Models
K. Eykholt et al. CVPR 2018.



Des modifications d'attitude et de position produisent des erreurs de classification.

Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al. CVPR 2019

Produis une image de la savane africaine ne montrant aucune girafe, montrant un lion, ne montrant aucune girafe, ne montrant absolument aucune girafe, ne montrant absolument aucun éléphant.

RA

You

Produis une image de la savane africaine ne montrant aucune girafe, montrant un lion, ne montrant aucune girafe, ne montrant absolument aucune girafe, ne montrant absolument aucun éléphant.



ChatGPT



L'image de la savane africaine a été créée en mettant l'accent sur un lion majestueux au centre, tout en respectant votre demande de ne pas inclure de girafes ni d'éléphants.



Message ChatGPT...



Generation de fausses images et désinformation



Questions éthiques et sociétales

Ethique et morale

- Ethique : du grec « èthos » ἦθος (caractère, coutume, mœurs)
- Morale : du latin « moralis » (relatif aux mœurs)
-> *Comment bien vivre ensemble en société?*



Paul Ricoeur (1913-2005)

Paul Ricoeur :

*« C'est par convention que je réserverai le terme d' « **éthique** » pour la visée d'une vie accomplie sous le signe des actions estimées bonnes et celui de « **morale** » pour le côté **obligatoire** marqué par des normes, des obligations, des interdictions caractérisées à la fois par une exigence d'universalité et par un effet de contrainte ».* (Ethique et morale 1990)

- L'Ethique est une réflexion individuelle ou collective sur les actes en fonction de valeurs
- La réflexion porte sur des choix dans un contexte donné compte-tenu de tensions entre les valeurs pour agir au mieux
- L'Ethique n'est pas la conformité à des normes ou des prescriptions ni à une législation

L'éthique

« *Que dois-je faire?* » (Kant)

Quelle est la décision, le choix, l'action la plus juste, compte-tenu de possibles tensions entre les **valeurs** humaines?

Décisions ayant un lien avec ce qui est bien et ce qui est mal

Réflexion contextuelle, dans une situation donnée

l'Éthique n'est pas la conformité à des normes ou des prescriptions ni à une législation

Exemple de valeurs

- Dignité
- Intégrité
- Bonheur
- Bien-être physique
- Bien-être mental
- Honneur
- Intimité
- Vie privée
- Amitié
- Vérité
- Liberté
- Egalité
- Equité
- Fraternité
- Démocratie
- Sécurité
- Soutenabilité

Valeurs intrinsèques, extrinsèques
Tension entre les valeurs

Reflexion éthique sur la technologie

- Effets sur les modes de vie, les valeurs humaines, la société
- Effets globaux et à long terme
- Effets cumulatifs
- Effets irréversibles
- Risques existentiels



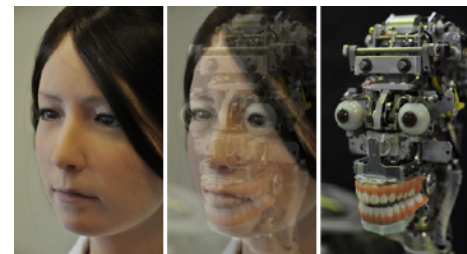
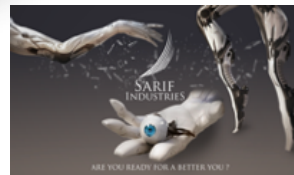
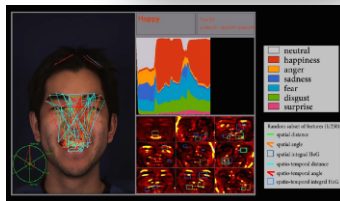
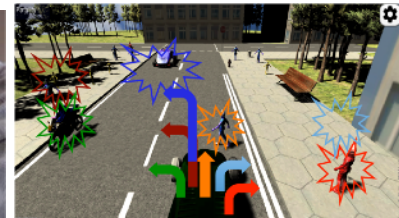
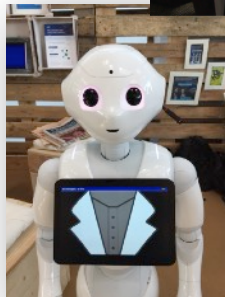
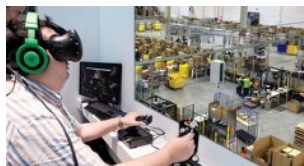
Quelques enjeux éthiques, sociétaux et juridiques de l'IA et de la Robotique

- Transformation du travail
- Vie privée, données personnelles
- Surveillance et notation des personnes
- Fausses informations et manipulation
- Biais des données

- Décisions de la machine et responsabilité
- Champ de la décision humaine
- Liens affectifs, attachement

- Transformation de l'être humain
- Imitation de l'identité humaine
- Statut du robot dans la société
- IA générale

- Certains usages et applications



Journal of Personality and Social Psychology
2016, Vol. 114, No. 2, 218–227

© 2016 American Psychological Association
1073-0909/16/\$12.00 http://dx.doi.org/10.1037/xap0000086

INNOVATIONS IN SOCIAL PSYCHOLOGY

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

Yilun Wang and Michal Kosinski
Stanford University



Théories éthiques

Les Grecs - Platon Aristote ...

- Ethique de la vertu: promeut des qualités morales de la personne et la « vie bonne »



Emmanuel Kant 1785

- C'est l'autonomie la liberté de choix qui rend possible la moralité.
- L'impératif catégorique: Agis uniquement d'après la maxime qui fait que tu peux vouloir en même temps qu'elle devienne une loi universelle.
- Ne jamais agir de telle sorte que nous traitions l'humanité en nous-mêmes ou chez les autres comme seulement un moyen mais toujours comme une fin en soi.



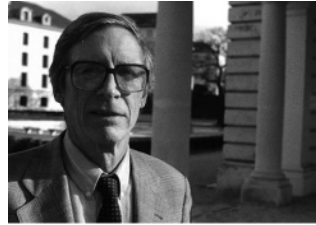
Jeremy Bentham 1789 John Stuart Mill 1861

- Conséquentialisme Utilitarisme: « Le plus grand bien pour le grand nombre »



- **Approche casuistique:** détermination au cas par cas

Théories éthiques modernes

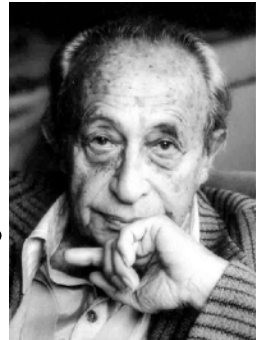


John Rawls 1971

- Théorie de la justice équité et protection du plus vulnérable (le voile de l'ignorance)

Hans Jonas 1979

- Le principe de responsabilité. Réflexion sur la portée des actions dans le temps et l'espace: *agis de façon que les effets de ton action soient compatibles avec la permanence d'une vie authentiquement humaine sur terre*

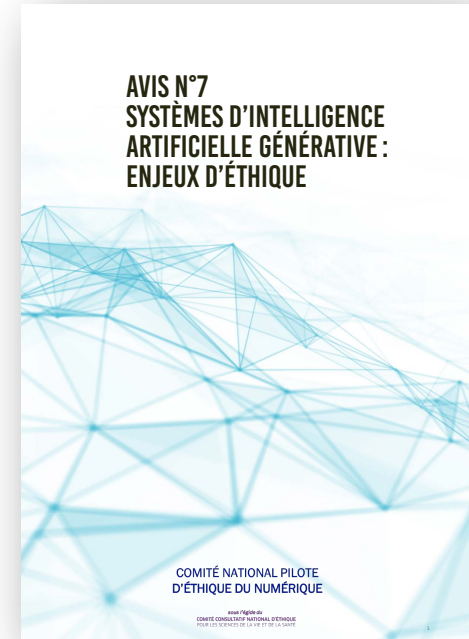


Limites de l'IA et de l'IA générative

- Opacité (milliards de paramètres); Explicabilité
- Pas de processus de validation et de vérification rigoureux
- Biais: qualité et représentativité des données; choix de conception
- Corrélations inappropriées ou fallacieuses
- Absence de lien causal entre données et résultat
- Absence de sémantique; pas d'inférence logique (ni probabiliste)

Enjeux éthiques et sociétaux de l'IA et de l'IA générative

- Question de la vérité; diffusion de fausses informations
- Discrimination
- Anthropomorphisation
- Comportements émergents
- Diversité culturelle et souveraineté
- Enjeux juridiques (PI, ...)
- Impacts environnementaux
- Impact sur les valeurs éducatives
- Impact sur les métiers et les emplois
- Responsabilité des fournisseurs de modèles/de systèmes/des utilisateurs



Impacts Environnementaux de l'IA

ESTIMATING THE CARBON FOOTPRINT OF BLOOM, A 176B PARAMETER LANGUAGE MODEL

Alexandra Sasha Luccioni
Hugging Face
sasha.luccioni@hf.co





Sylvain Viguier
Graphcore
sylvainv@graphcore.ai

Anne-Laure Ligozat
LISN & ENSIIE
anne-laure.ligozat@lisn.upsaclay.fr



oecd.ai/footprint

Direct environmental impacts AI compute resources lifecycle

Production 	Transport 	Operations 	End-of-life 
<ul style="list-style-type: none"> Raw material extraction Assembly Manufacturing 	<ul style="list-style-type: none"> Distribution Freight transportation Handling & storage 	<ul style="list-style-type: none"> Energy consumption Water consumption Carbon footprint 	<ul style="list-style-type: none"> Collection & shipping Dismantling & recycling Waste disposal

Indirect environmental impacts AI compute applications

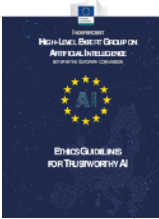
Positive impacts	Negative impacts
<ul style="list-style-type: none"> Beneficial sectoral applications Climate mitigation and adaptation Environmental modelling and forecasting 	<ul style="list-style-type: none"> Harmful sectoral applications Carbon leakage (net increase in emissions) Consumption patterns and rebound effects

Principes d'éthique biomédicale

- **Principe d'autonomie:** obligation de respecter les capacités de décisions et le consentement des personnes autonomes;
- **Principe de bienfaisance:** obligation de procurer des bénéfices et de mesurer les bénéfices par rapport aux risques;
- **Principe de non-malfaisance:** obligation d'éviter de nuire;
- **Principe de Justice:** obligation d'équité, juste distribution des bénéfices et des risques.

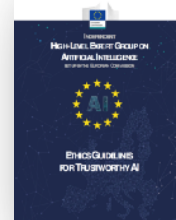
Principes d'éthique du numérique

- **Principe d'autonomie:** préserver l'agentivité et le contrôle humains;
- **Principe de non-malfaisance:** ne pas nuire ni exacerber un mal (sûreté, sécurité, robustesse technique);
- **Principe de Justice:** équité, réduction du biais, non discrimination, proportionnalité ;
- **Principe d'explicabilité:** transparence, interprétabilité, traçabilité, auditabilité.



Exigences pour une IA de confiance

Groupe d'Experts de haut niveau sur l'IA (UE) Avril 2019

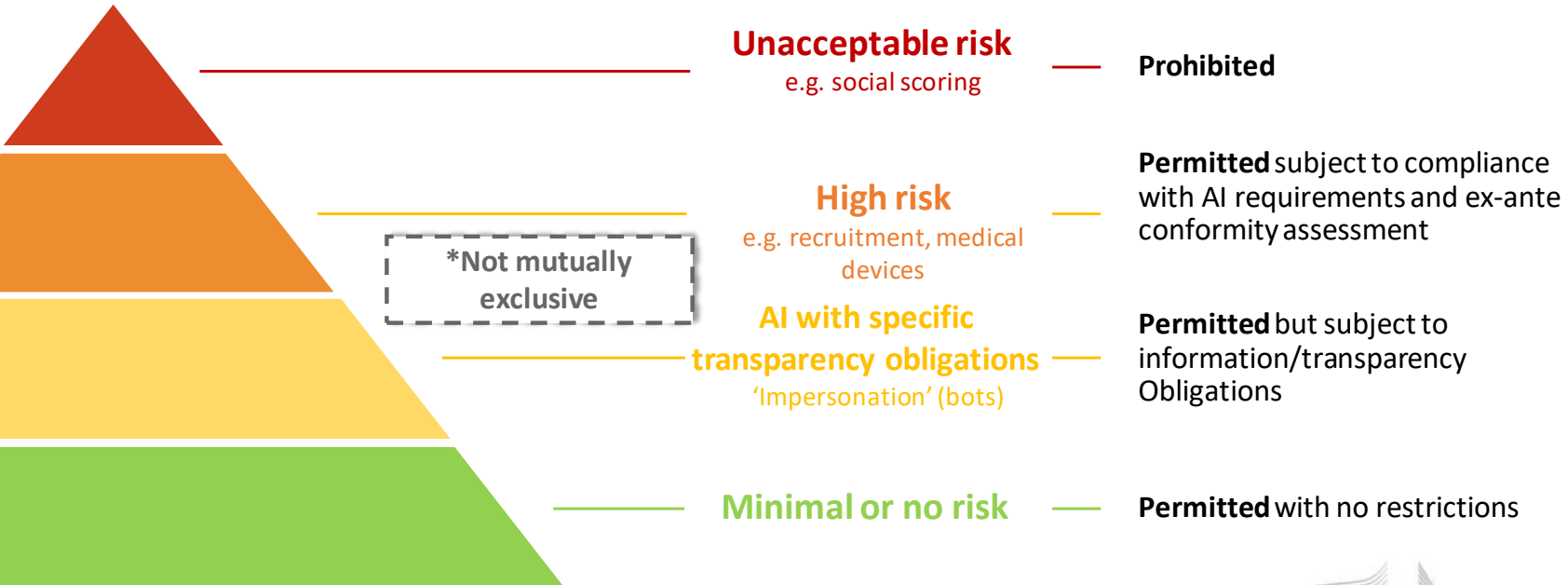


1. **Action et supervision humaine**
2. **Robustesse technique et sécurité** : résilience et sûreté de fonctionnement, sécurité, précision, fiabilité, reproductibilité
3. **Respect de la vie privée et gouvernance des données**: qualité et intégrité des données
4. **Transparence**: traçabilité, explicabilité, communication
5. **Diversité non-discrimination et équité**: absence de biais injustes, accessibilité, participation des parties prenantes
6. **Bien-être social et environnemental**: durabilité, respect de l'environnement, impact social, démocratie
7. **Responsabilité** : auditabilité, réduction des incidences négatives, communication, arbitrages, recours.

<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>

Réglementation européenne sur l'IA (AI Act)

Une approche basée sur le risque présenté par l'usage auquel est destiné le système



En guise de conclusion

L'IA et l'Apprentissage machine en particulier sont des méthodes et des techniques efficaces pour le traitement des données et l'interprétation automatique

Les systèmes basés sur l'apprentissage machine doivent être rendus plus robustes et plus stables

L'IA n'est pas la solution miracle à tous les problèmes et dans toutes les domaines

L'utilisation de l'IA dans des domaines relevant des droits humains doit être abordée avec beaucoup de précautions (voire ne pas être utilisée)

La formation à l'IA doit aussi traiter de ses limites

L'explicabilité et la transparence sont des qualités nécessaires pour la confiance dans la technologie

Les contextes à haut risque nécessitent la mise en place de mécanismes de vérification et de validation solides et une gouvernance appropriée