Le champ de l'*Explainable AI* :

Des machines pour expliquer les machines ?

Nicolas Berkouk (EPFL) - Conférence IA et Santé : Approches inter-disciplinaires, Nantes 01/07/2022

Structure de la présentation

(1) Expliquer les réseaux de neurones : pourquoi ?

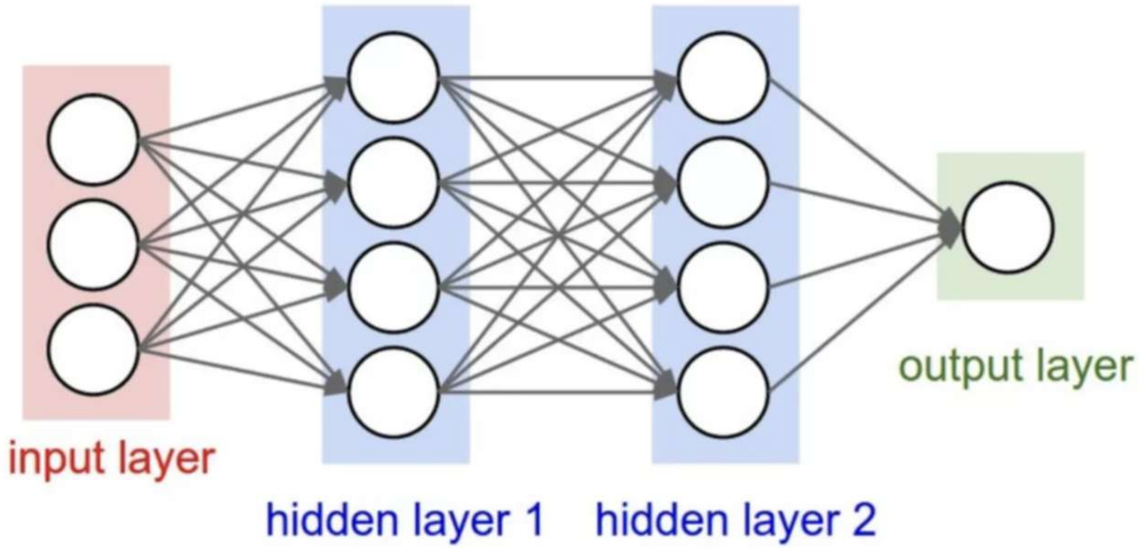
(2) Le champ de l'*Explainable AI* : genèse et propriétés

(3) La mise en technique de la production d'explications : quelles conséquences, quels problèmes ?

Expliquer les réseaux de neurones : pourquoi ?

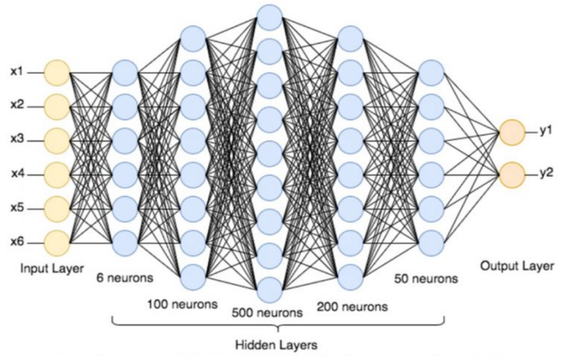
Réseaux de neurones : de quoi parle-t-on ?

- Classe de fonctions, représentées par un graphe de calcul, dont les poids des arrêtes sont les paramètres
- Pour lesquelles il est « facile » de calculer la « dérivée » (gradient)

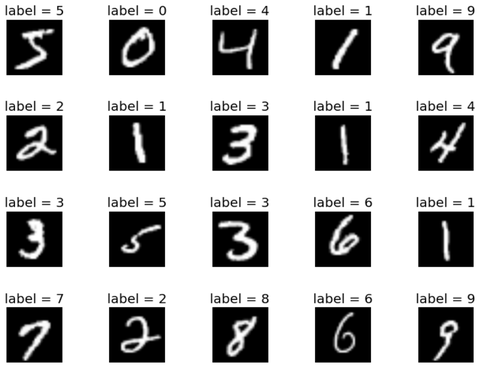


Expliquer les réseaux de neurones : pourquoi ?

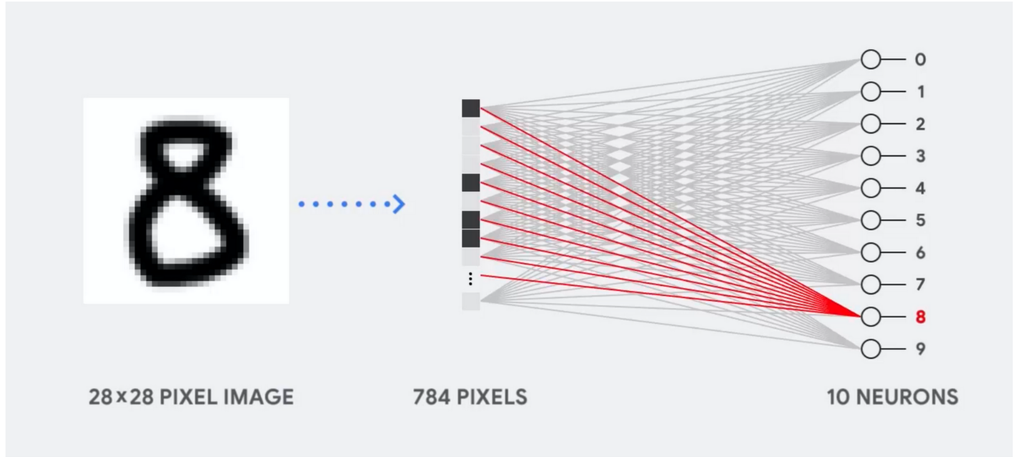
Comment utilise-t-on un réseau de neurone ? (Apprentissage supervisé)



(1) Choix de l'architecture du modèle



(2) Entraînement sur une base de donnée labellisée



(3) Utilisation sur des données

Expliquer les réseaux de neurones : pourquoi ?

Symbolisme vs. Connexionnisme : la controverse structurelle de l'IA [Cardon *et. al.* 2018]

Deux ensembles de méthodes d'apprentissage machine qui s'opposent *dans la pratique* parce qu'elles relèvent de deux *projets théoriques* radicalement différents.

- Paradigme symboliste : L'algorithme effectue une suite de calculs pré-définis à partir de variables qui ont un sens dans le monde du programmateur.
  - Systèmes experts, Régressions, Arbres de décision, *Handcrafted features* en vision…
  - Idéal positiviste de l'apprentissage machine.
  - Systèmes interprétables par construction (Bachimont, 1996)

- Paradigme connexionniste : L'algorithme prend « le monde » en entrée, et par un processus d'entraînement, optimise les calculs à effectuer pour être le plus performant sur une tâche.
  - Réseaux de neurones
  - Idéal d'un apprentissage machine agnostique
  - Interprétabilité / Explicabilité ?

Expliquer les réseaux de neurones : pourquoi ?

Une mise en oeuvre massive et en rapide expansion en dehors des laboratoires…

- Les réseaux de neurones : des résultats fulgurants,
- Particulièrement pour l'analyse de données non-structurées (images, texte, audio),
- Qui sont captées quotidiennement par différents médiums (smartphone, traces de navigation web, objets connectés…).
- Nous sommes d'ores et déjà quotidiennement en interaction avec des réseaux de neurones,
- Et pourtant seulement au tout début de l'histoire de leurs applications : systèmes de recommandation, médecine personnalisée, voiture autonome…

Expliquer les réseaux de neurones : pourquoi ?

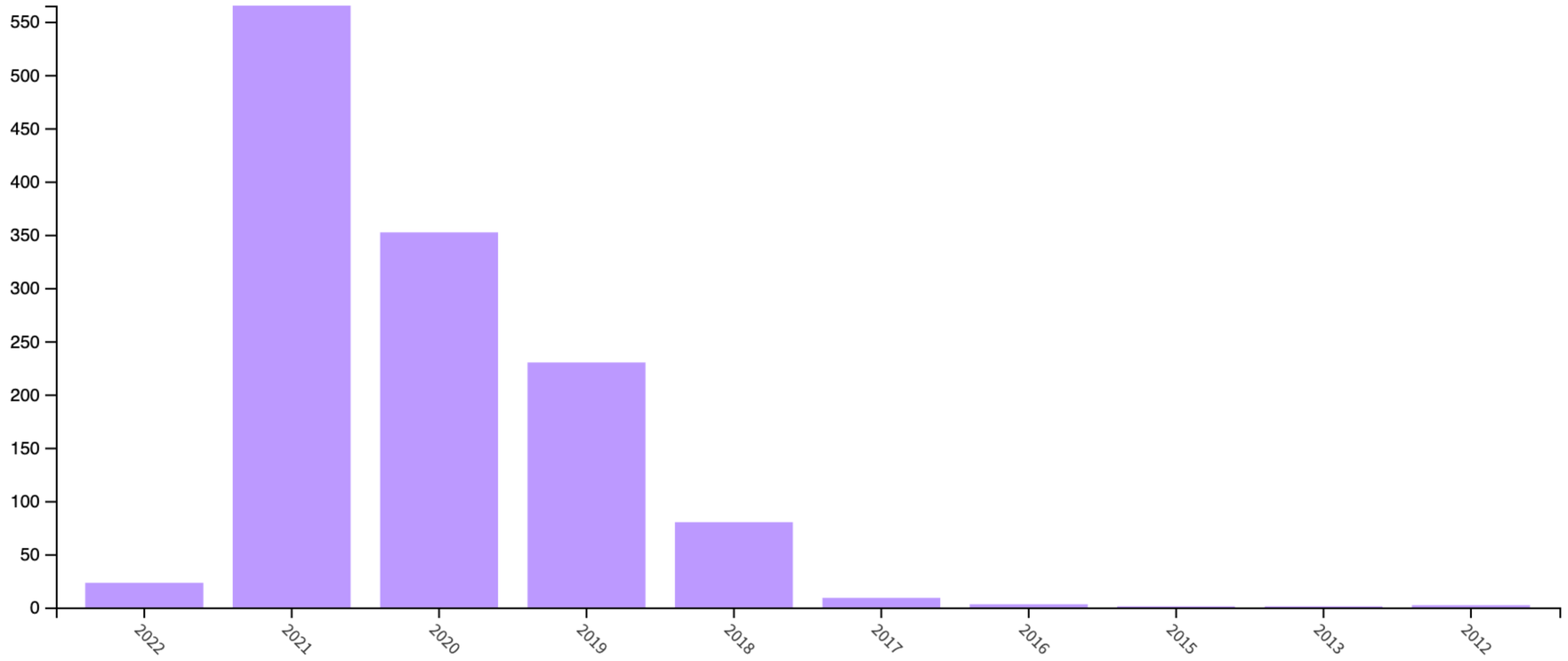
…Qui fonde le besoin d'intelligibilité des résultats de ces algorithmes sur plusieurs dimensions

- Légale (RGPD) : droit individuel à l'information des usagers concernés par des traitements algorithmiques mise en oeuvre par des administrations.

- Économique : 90% des entreprises utilisant des réseaux de neurones considèrent comme critique la capacité à expliquer comment un algorithme a aboutit à une décision (Source : *Global AI Adoption Index*, IBM, 2021).

Le champ de l'*Explainable AI* : genèse et propriétés

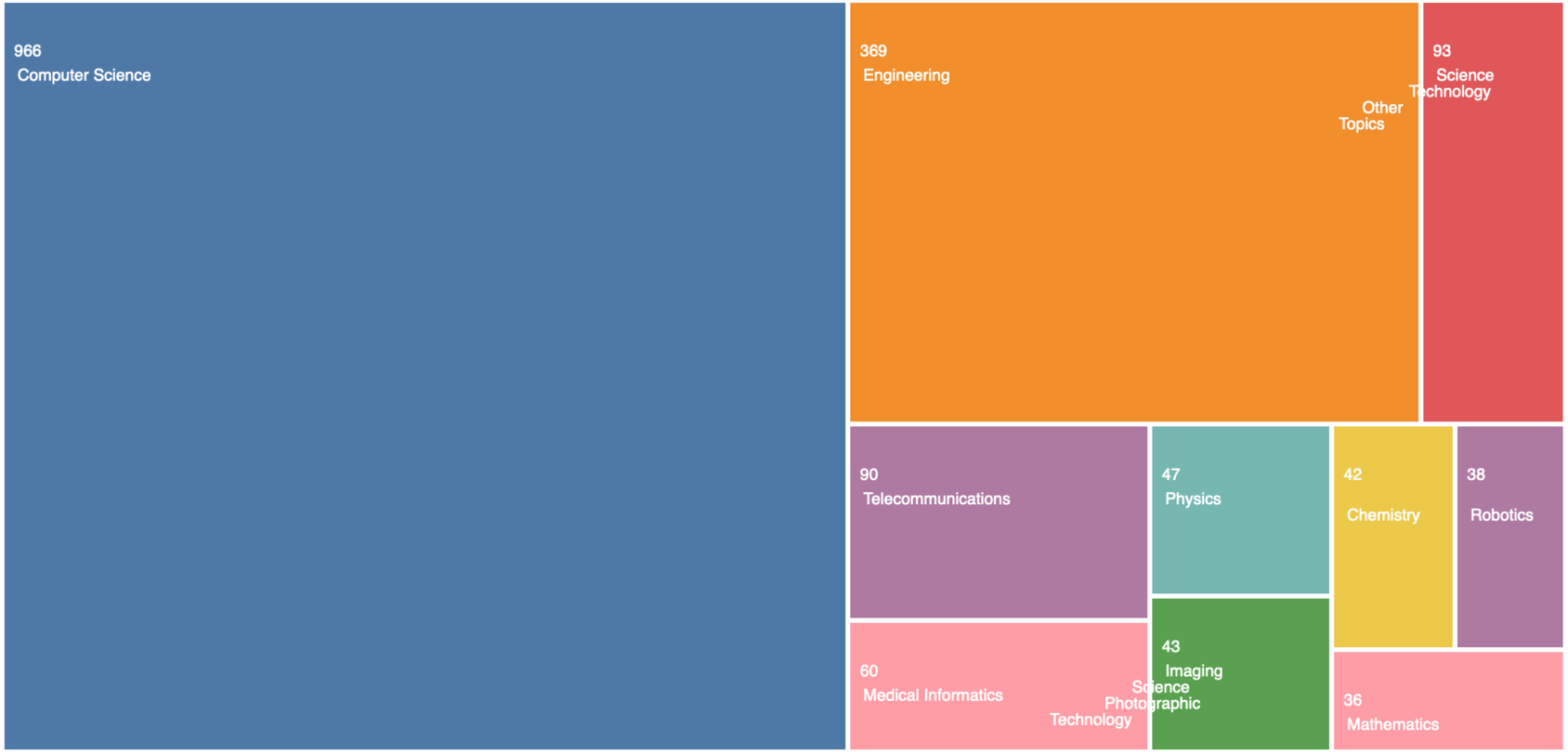
Soudaineté



Nombre d'articles scientifiques par année contenant 'Explainable AI' - Source : *Web of Science*

Le champ de l'*Explainable AI* : genèse et propriétés

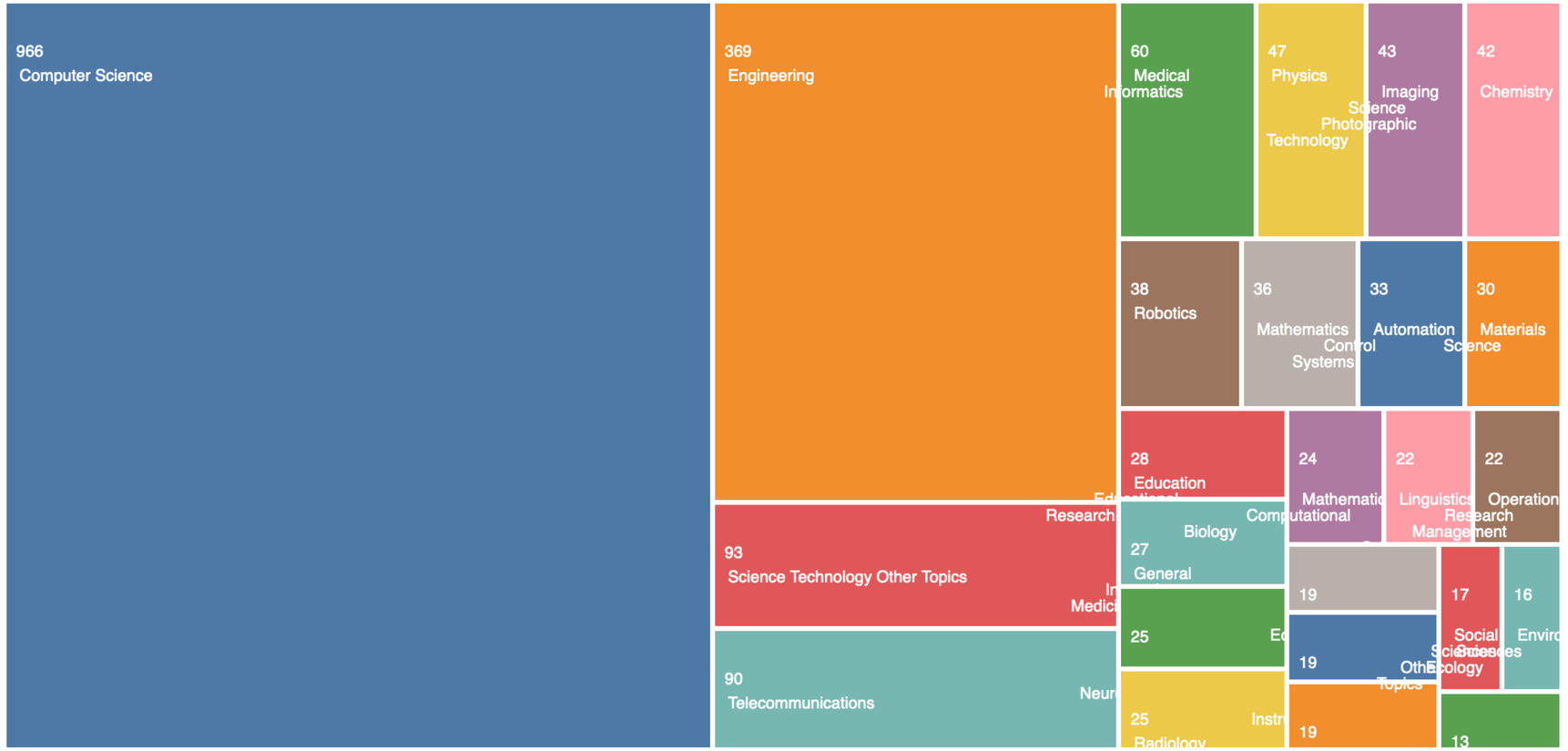
Appartenance disciplinaire



Discipline des articles scientifiques contenant 'Explainable AI' - Source : *Web of Science* (catégories de *Web of Science)*

Le champ de l'*Explainable AI* : genèse et propriétés

Appartenance disciplinaire

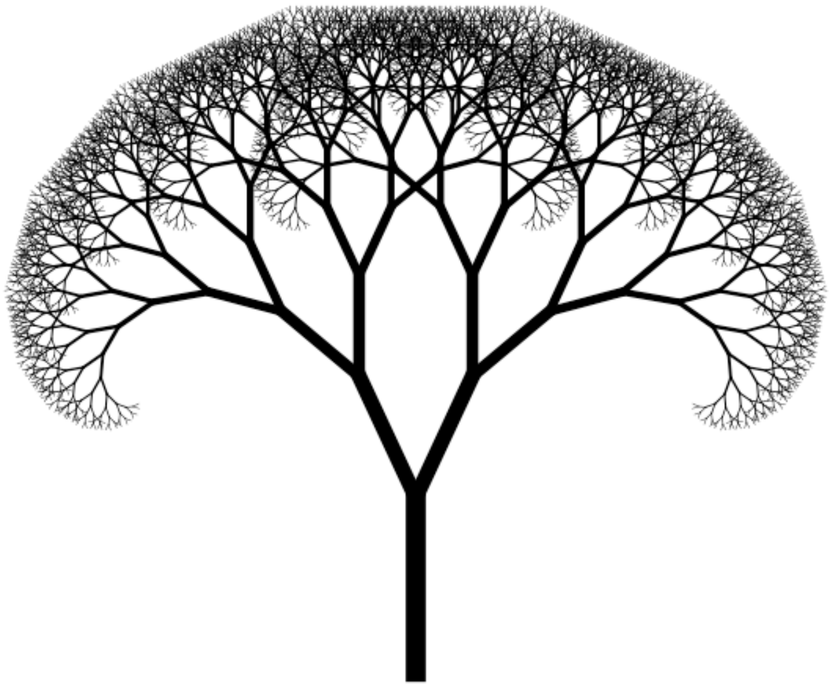


Discipline des articles scientifiques contenant 'Explainable AI' - Source : *Web of Science* (catégories de *Web of Science)*

Le champ de l'*Explainable AI* : genèse et propriétés

La mise en technique de la production d'explication : deux hypothèses

- Division fractale du travail scientifique contemporain (*Chaos of Disciplines,* Abbott)



Le champ de l'*Explainable AI* : genèse et propriétés

La mise en technique de la production d'explication : deux hypothèses

- Intérêt économique et capture culturelle (Carpenter, Moss, Kwak)

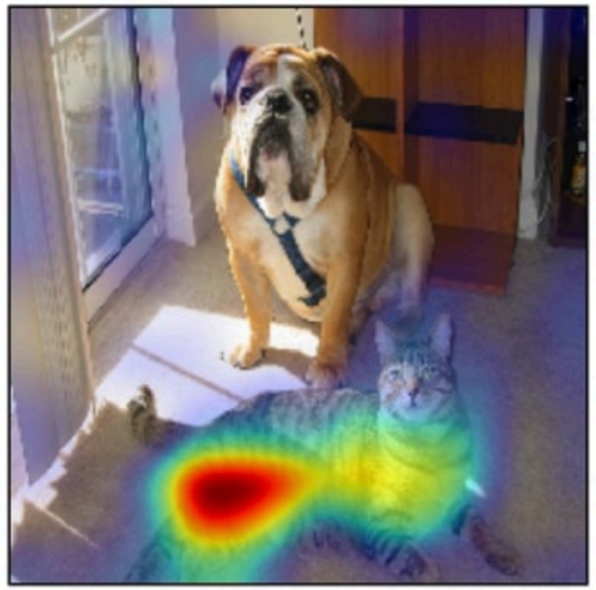
Most important aspects of AI trust and explainability[1]

90%	Maintaining brand integrity and customer trust
89%	Meeting external regulatory and compliance obligations
89%	Meeting internal reporting obligations
88%	Ability to monitor and govern data and AI across its lifecycle
87%	Ensuring applications and services minimize bias

Source : *Global AI Adoption Index*, IBM, 2021

La mise en technique de la production d'explication : quelles conséquences, quels problèmes?

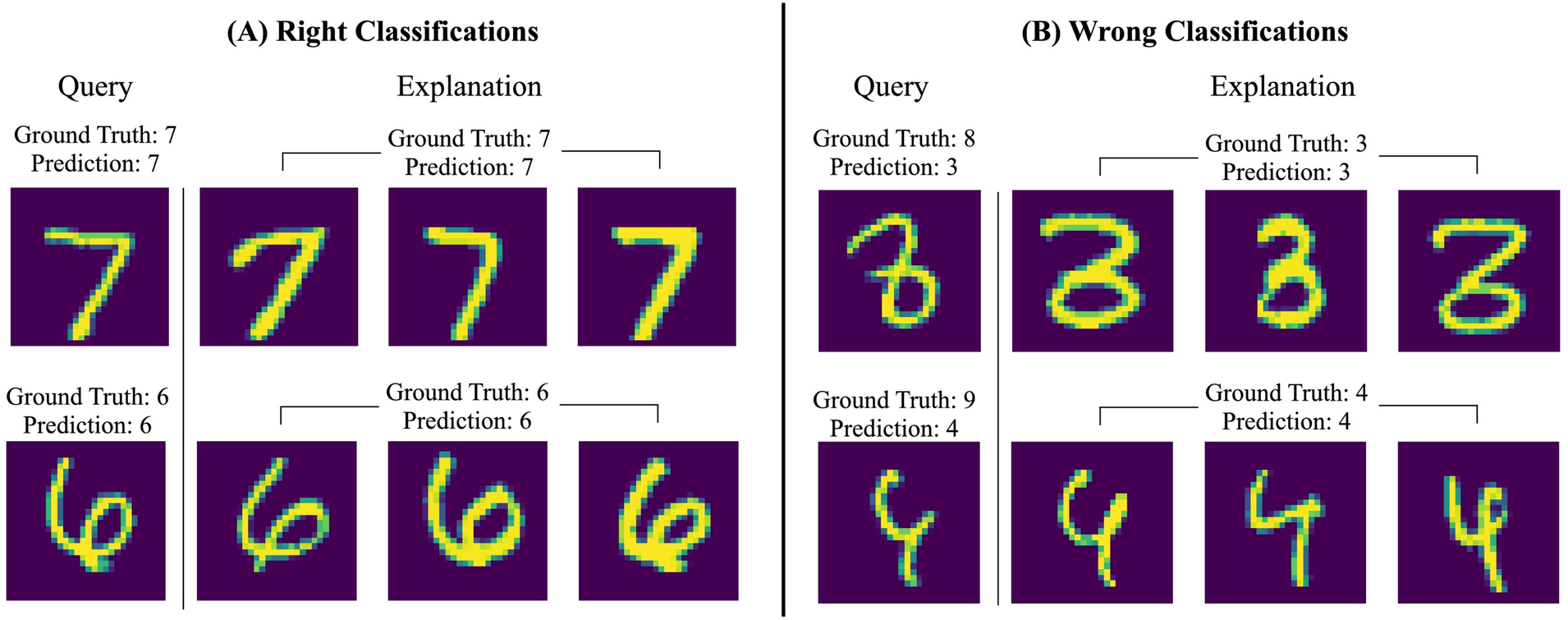
Hétérogénéité des conceptions d'explications : GradCAM



Source : *Grad-cam: Visual explanations from deep networks via gradient-based localization*, Selvaraju *et. al.*, International Journal of Computer Vision, 2019

La mise en technique de la production d'explication : quelles conséquences, quels problèmes?

Hétérogénéité des conceptions d'explication : Exemples *post-hoc*



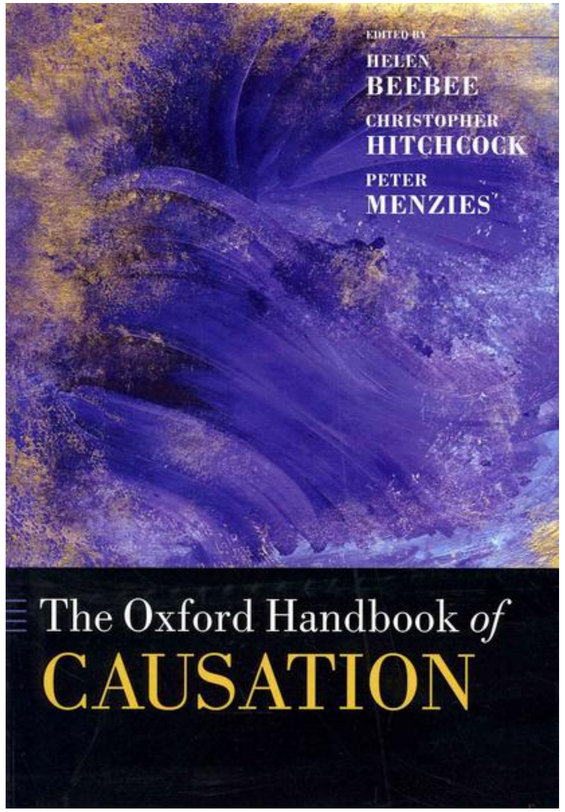
Source : *Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xAI user studies*, Kenny *et. al.*, Artificial Intelligence, 2021

La mise en technique de la production d'explication : quelles conséquences, quels problèmes?

Le regard de la philosophie de l'imputation de causes

- Il n'existe pas de théorie unifiée de la causation qui puisse satisfaire simultanément les exigences des programmes de recherche en éthique, épistémologie, ou en ontologie.

- Cette théorie unifiée ne verra probablement jamais le jour.



La mise en technique de la production d'explication : quelles conséquences, quels problèmes?

Quelques questions

- Il n'est pas raisonnable de vouloir formuler une théorie unifiée de l'explicabilité des réseaux de neurones
- Dès lors, peut-on mettre en relation les différentes théories de l'imputation de causes et celles de l'explicabilité des réseaux de neurones ? En quoi diffèrent-elles ?
- Quel est le régime épistémologique de l'IA connexionniste ? Que peut-on savoir des réseaux de neurones ? Que peuvent-ils nous apprendre sur le monde (physique, biologique, sociale) ?
- Quelles sont les conditions de possibilité et les limitations des différentes conceptions techniques de l'explicabilité ?
- Existe-t-il une correspondance entre l'espace des positions (institutionnelles, nationale) des chercheurs en *Explainable AI* et les différents types d'explications proposées ?

Merci pour votre attention