Entrainer, évaluer et tester une Intelligence Artificielle fiable

Atelier technique

Sarah LEROY | Ingénieure développement et test de l'IA
Clément FRANCOIS | Ingénieur d'études et test IA
Avec le support de Anne-Laure WOZNIAK
13/06/2023

Kereval

Conseil
Audit de processus de tests

Innovation & projets R&D

Stratégie de tests et conception

Développement & maintenance de bancs de tests

Exécution de tests

Automatisation & Devops

Organisation de Connectathons et Projectathons

NOS FORMATIONS

NOS LABELLISATIONS

Sommaire

- **Introduction**
  - Qu'est-ce que l'IA ?
  - Etapes d'un projet de Machine Learning
  - Différences entre le test logiciel et le test de l'IA
- **Evaluer un modèle IA**
  - Différences entre l'évaluation et le test
  - Métriques d'évaluation
- **Live Coding : Données, entrainement et évaluation**
  - Préparation des données
  - Entrainement d'un réseau de neurones
  - Evaluation du modèle
- **Tester un modèle IA**
  - Importance & défis du test de l'IA
  - Propriétés des modèles de ML
    - Qualité des données
    - Ethique & équité
    - Explicabilité
    - Robustesse
- **Live Coding : Test**
  - Robustesse
  - Explicabilité
  - Ethique

- Qu’est-ce que l’IA ?
- Etapes d’un projet de Machine Learning
- Différences entre le test logiciel et le test de l’IA

Introduction

Qu’est-ce que l’IA ?

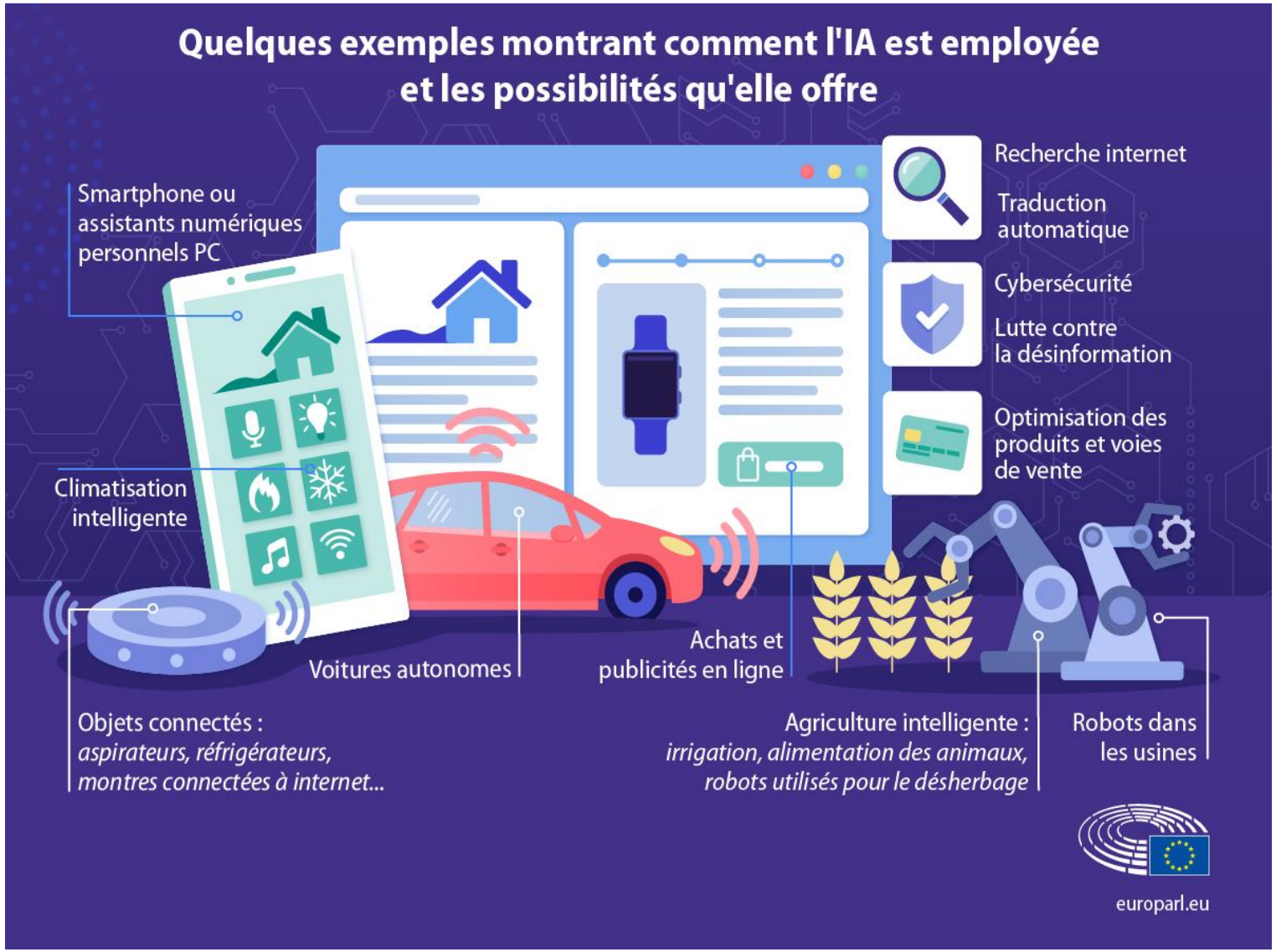
« Possibilité pour une machine de **reproduire des comportements liés aux humains**, tels que le **raisonnement**, la **planification** et la **créativité**. »

Parlement européen

« Capacité d’un système technique à **acquérir, traiter et appliquer des connaissances et des compétences**. »

ISO/IEC TR 29119-11:2020

Utilisation actuelle



Narrow AI

(ou IA étroite, IA faible)

Système programmé pour effectuer une tâche spécifique, dans un contexte limité.

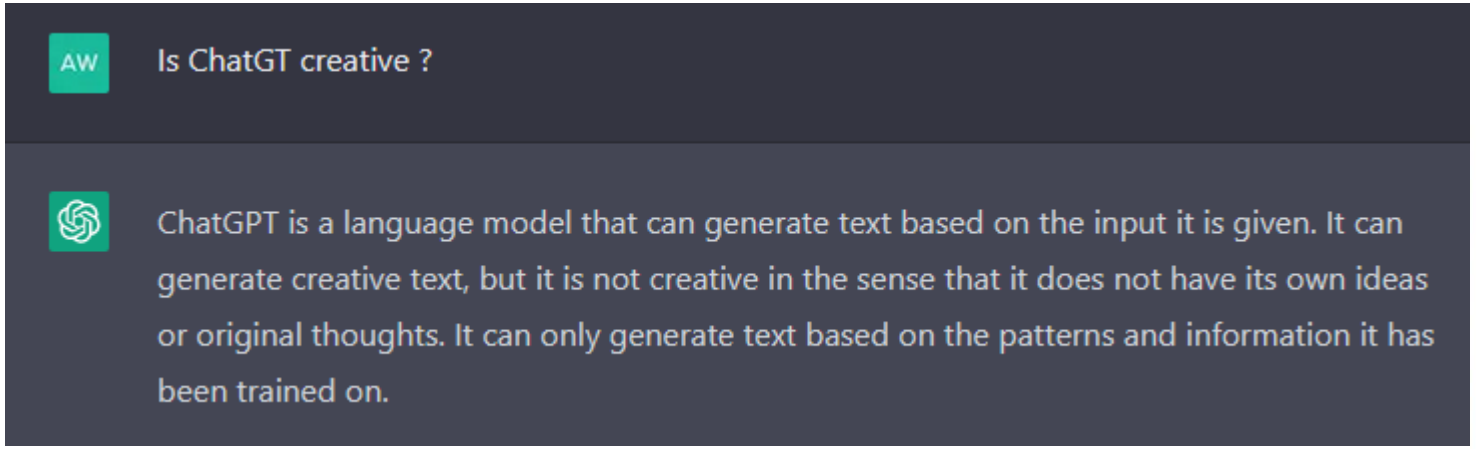
Source : [1]

Vers une IA générale

Système qui possède des **capacités cognitives générales, similaires à celles des humains**.

Quid de la **créativité** ?

- Générateur d'images
- Robot conversationnel



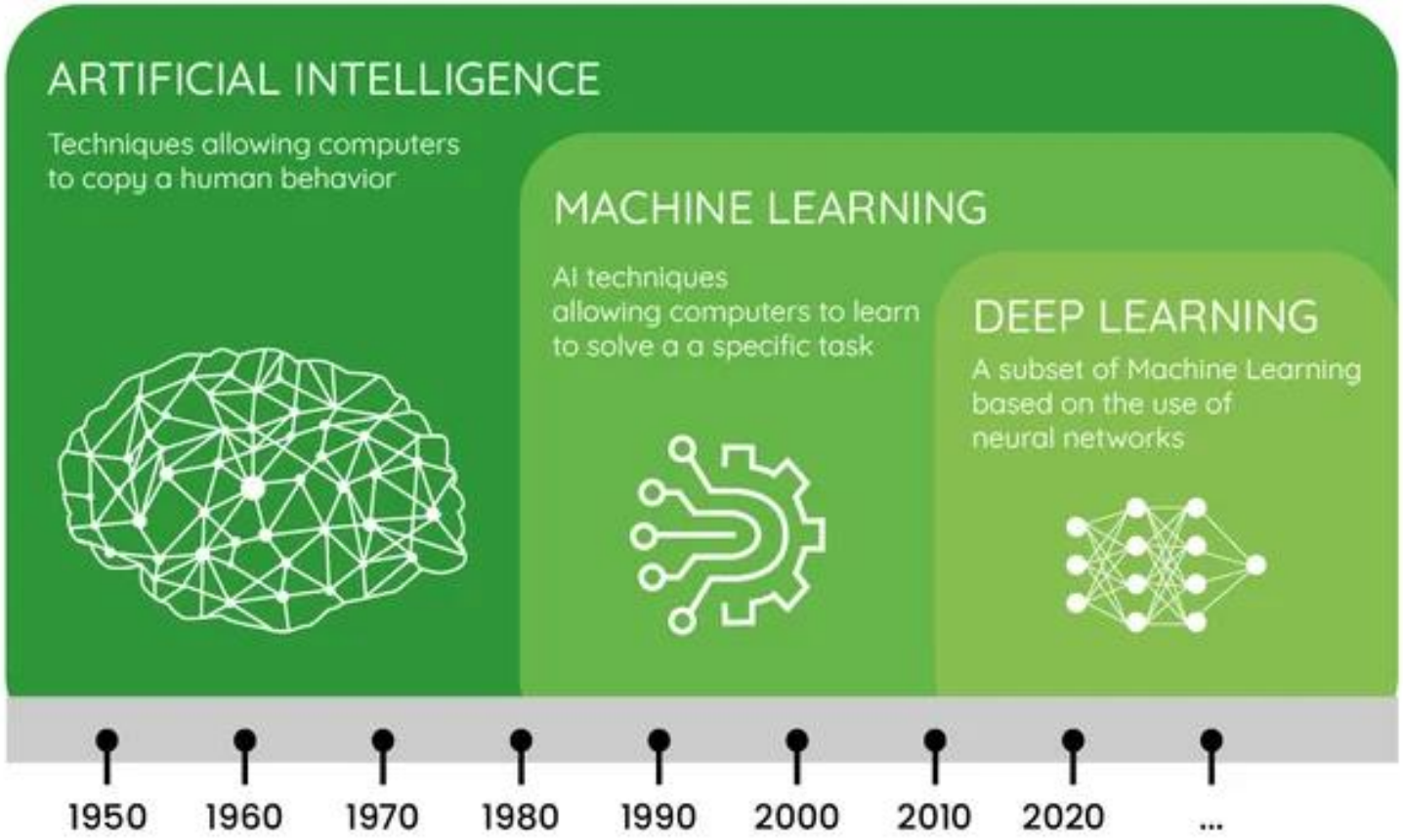
Source : [2]

Source : [3]

IA, Machine Learning et Deep Learning

Machine Learning (ML) : apprentissage automatique

Deep Learning (DL) : apprentissage profond

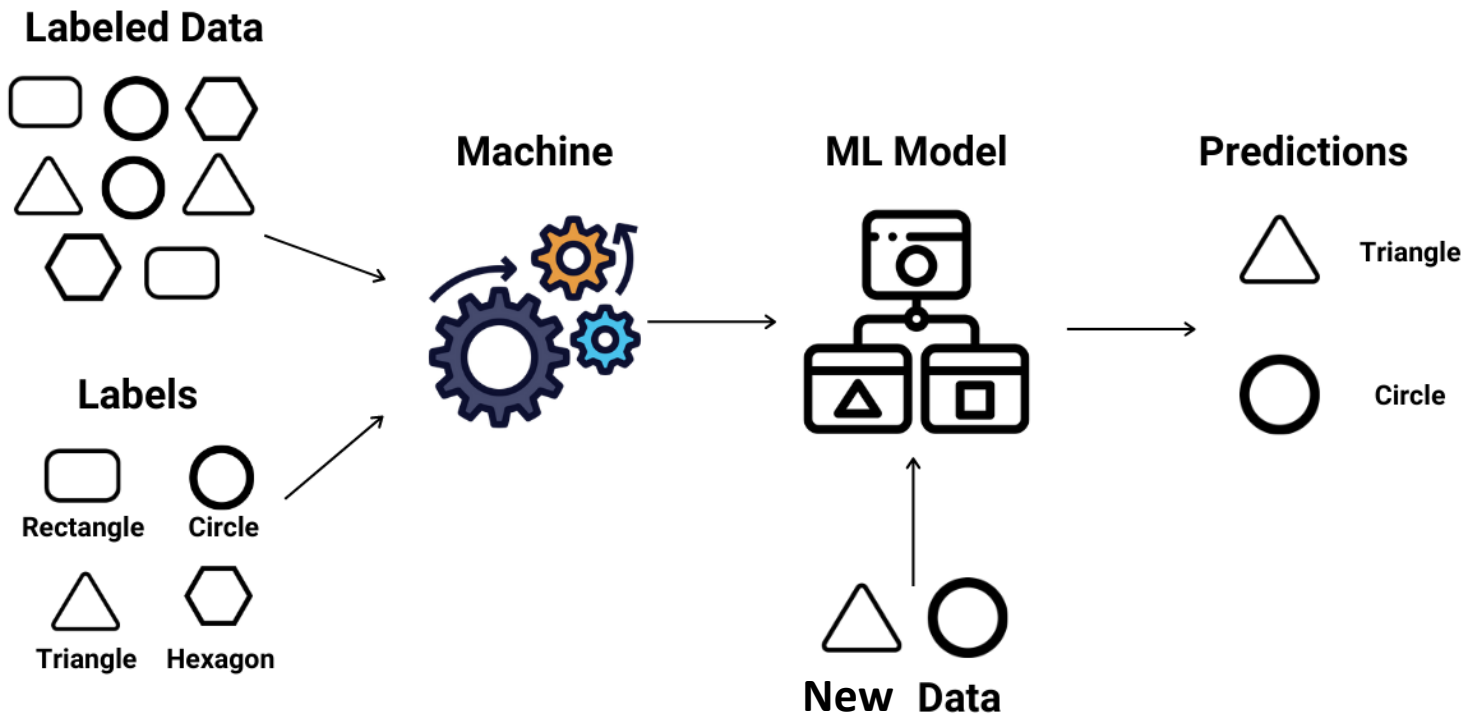


Source : [4]

Types de ML (1/3)

Apprentissage supervisé

Tâches réalisées : classification, régression

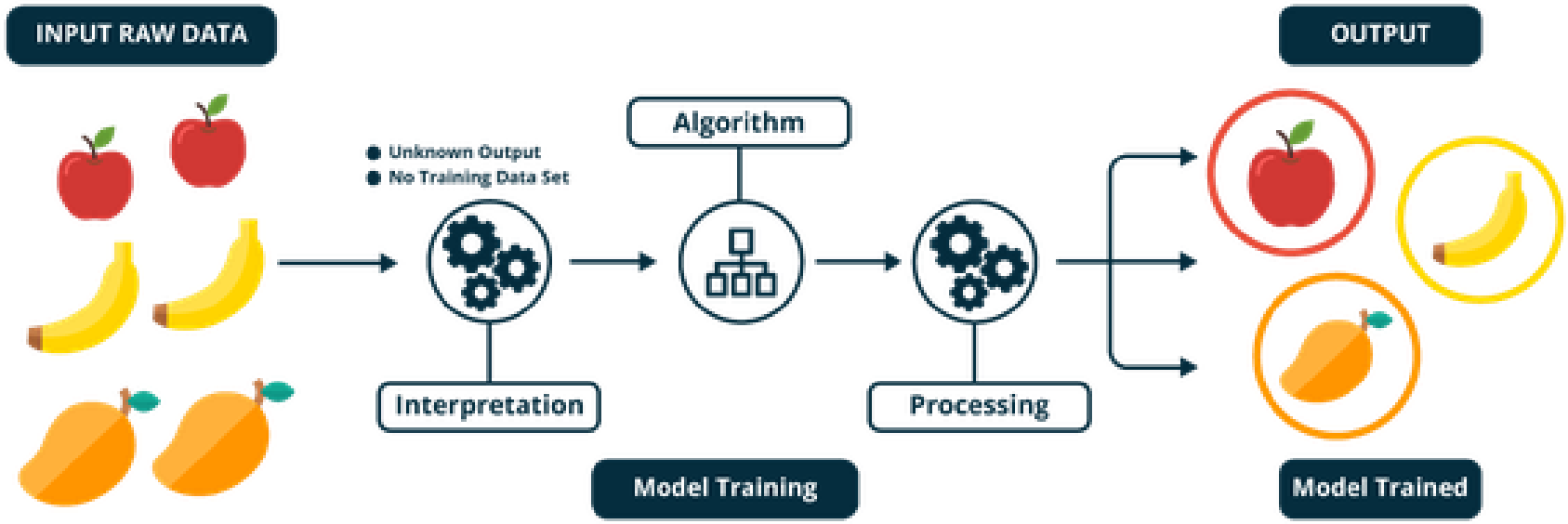


Source : [5]

Types de ML (2/3)

Apprentissage non-supervisé

Tâches réalisées : association, regroupement (clustering)

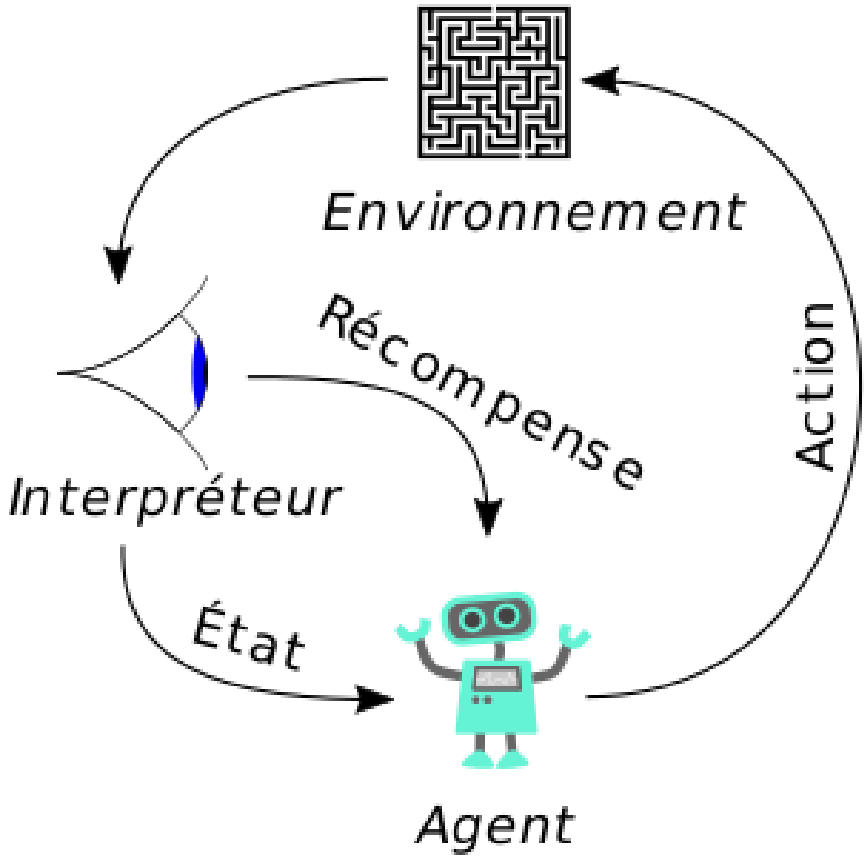


Source : [5]

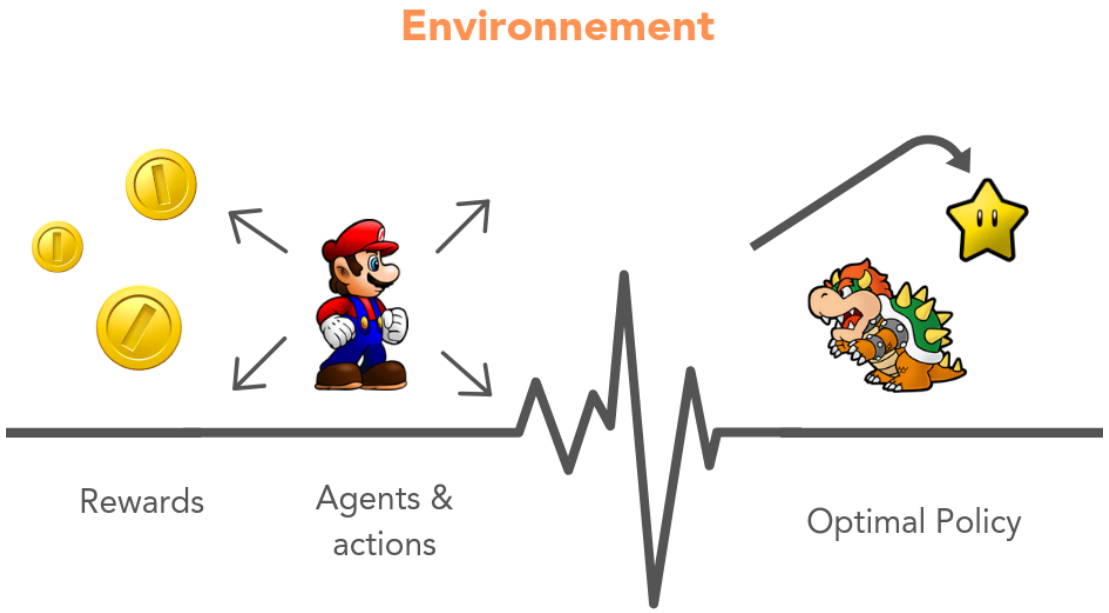
Types de ML (3/3)

Apprentissage par renforcement

Tâches réalisées : interactions avec l'environnement, prise de décision



Source : [5]



Source : [6]

Etapes d'un projet de Machine Learning (1/3)

« **Processus** utilisant des techniques informatiques pour permettre aux systèmes **d'apprendre à partir de données ou d'expériences** »

ISO/IEC TR 29119-11

LES 7 ÉTAPES DU **MACHINE LEARNING**

Algorithme (d'apprentissage) : programme qui produit un modèle ML à partir d'un ensemble de données d'apprentissage.

Modèle : sortie du processus d'apprentissage, objet informatique qui associe des données d'entrée à une sortie.

Hyperparamètre : paramètre utilisé pour contrôler l'apprentissage d'un modèle ML ou pour définir sa configuration, son architecture.

Etapes d'un projet Machine Learning (2/3)

LES 7 ÉTAPES DU **MACHINE LEARNING**

Collecte et préparation des données

Activités d'acquisition des données, de prétraitement des données et d'ingénierie des caractéristiques.

Activités de **nettoyage**, de **transformation**, d'**augmentation** et d'échantillonnage des données.

Activité d'**identification des attributs** des données brutes les plus pertinents.

Séparation des données en trois sous-ensembles : données d'entraînement, données de validation et données de test.

Etapes d'un projet de Machine Learning (3/3)

Entraînement

Processus d'**application de l'algorithme ML** à l'ensemble de données d'apprentissage pour créer un modèle ML.

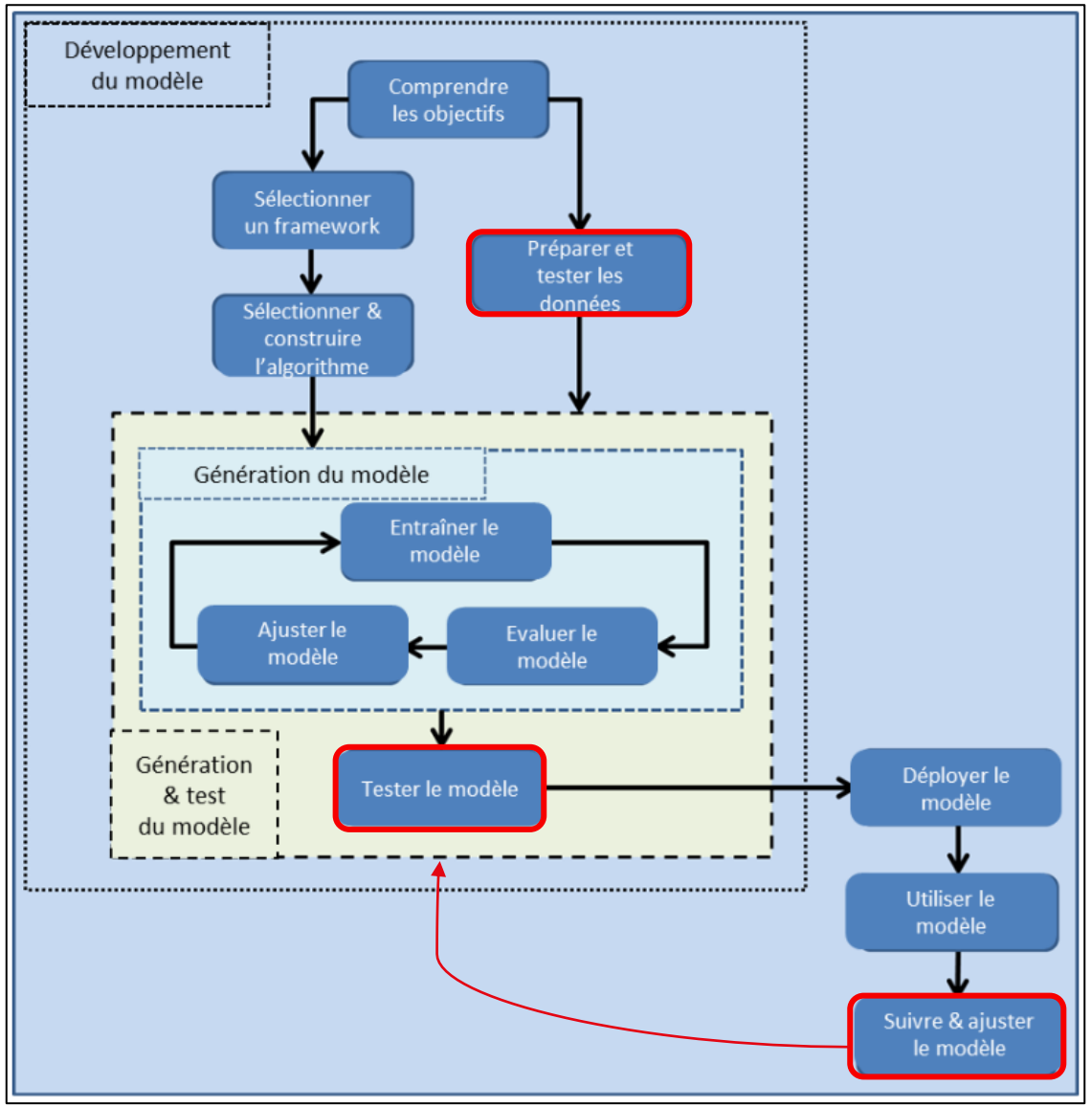
Evaluation

Processus de **comparaison des mesures de performance** aux critères requis et/ou à ceux d'autres modèles.

Ajustement

Processus permettant de **déterminer les hyperparamètres optimaux** en fonction d'objectifs de performances.

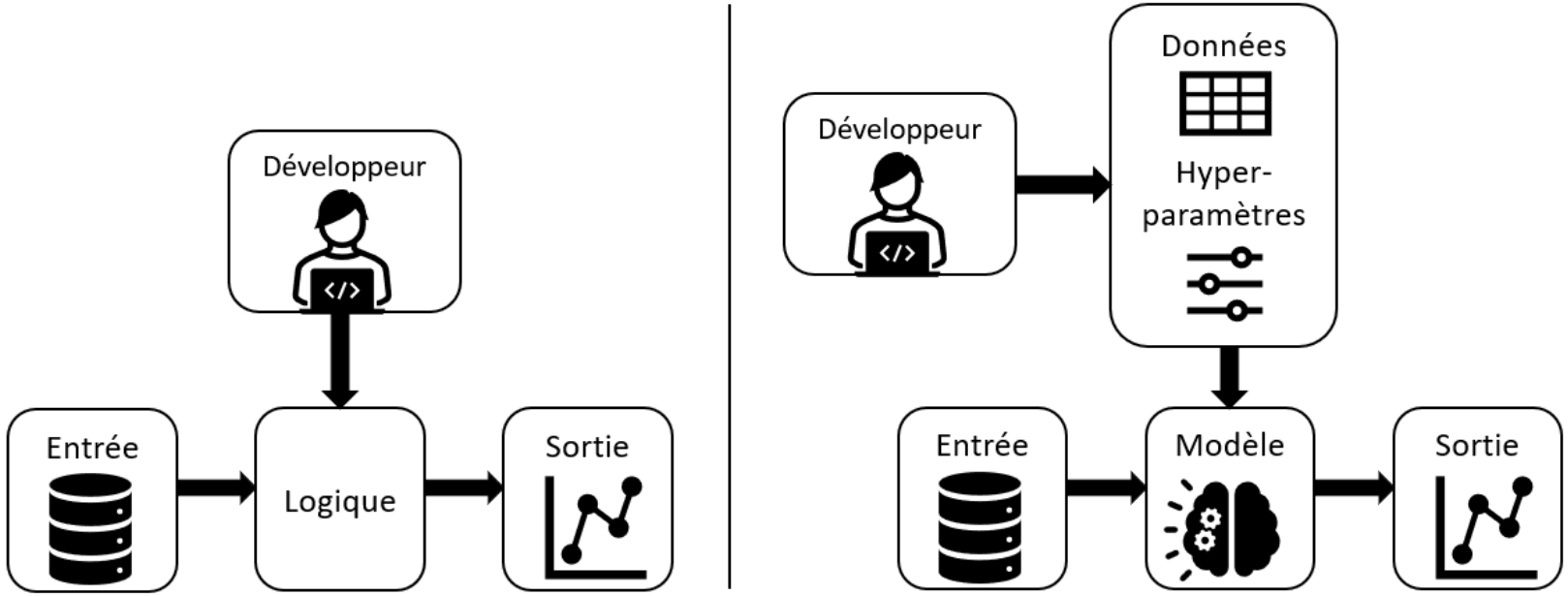
Processus du ML



Source : ISTQB CT-AI Syllabus v1.0 [7]

Test logiciel vs test de l'IA

- Le comportement du système n'est **pas programmé explicitement**
- La logique du programme est **déterminée par les données d'entrainement**
- Les processus d'entrainement sont **stochastiques**

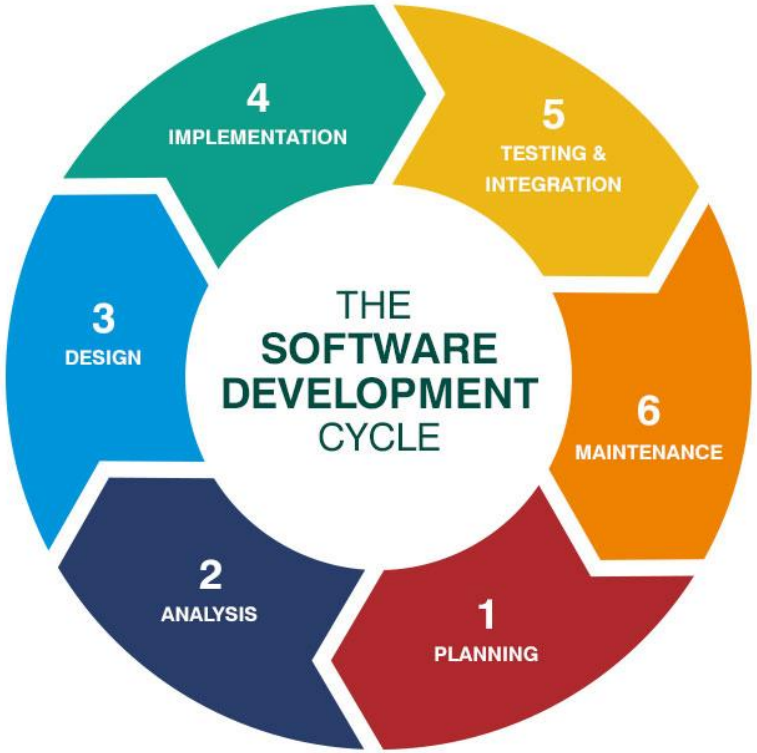


- Différences entre l’évaluation et le test
- Métriques d’évaluation

Evaluer un modèle

Différences entre l'évaluation et le test

LES 7 ÉTAPES DU **MACHINE LEARNING**



Source : [8]

Test ou évaluation ?

L'évaluation du modèle couvre l'ensemble des métriques et graphiques qui résument la performance sur un jeu de données de validation ou de test.

Le test du modèle implique des vérifications explicites du comportement attendu de la part du modèle.

Métriques d'évaluation – Classification binaire (1/2)

Modèles de classification supervisés :

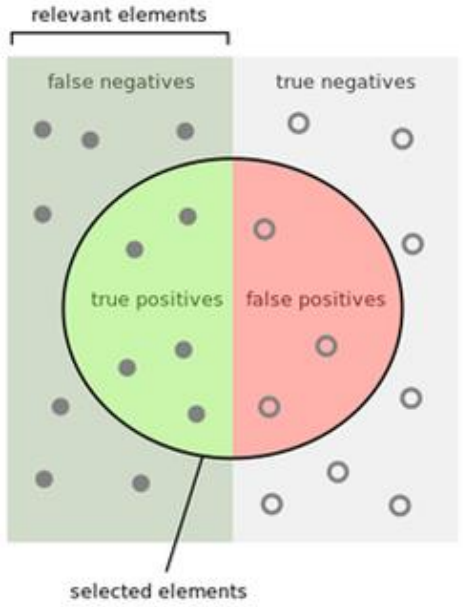
- Matrice de confusion

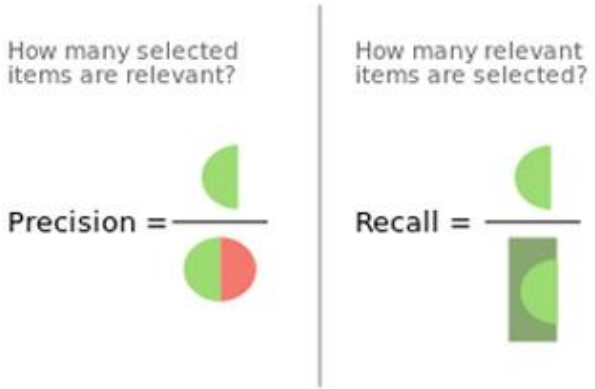
		Réel	
		Positif	Négatif
Prédit	Positif	VP	FP
	Négatif	FN	VN

- **Accuracy**

$$\frac{VP + VN}{VP + VN + FP + FN}$$

- Precision : $\frac{VP}{VP + FP}$
- Rappel (Recall) : $\frac{VP}{VP + FN}$
- Score F1 : $2 * \frac{Precision * Rappel}{Précision + Rappel}$



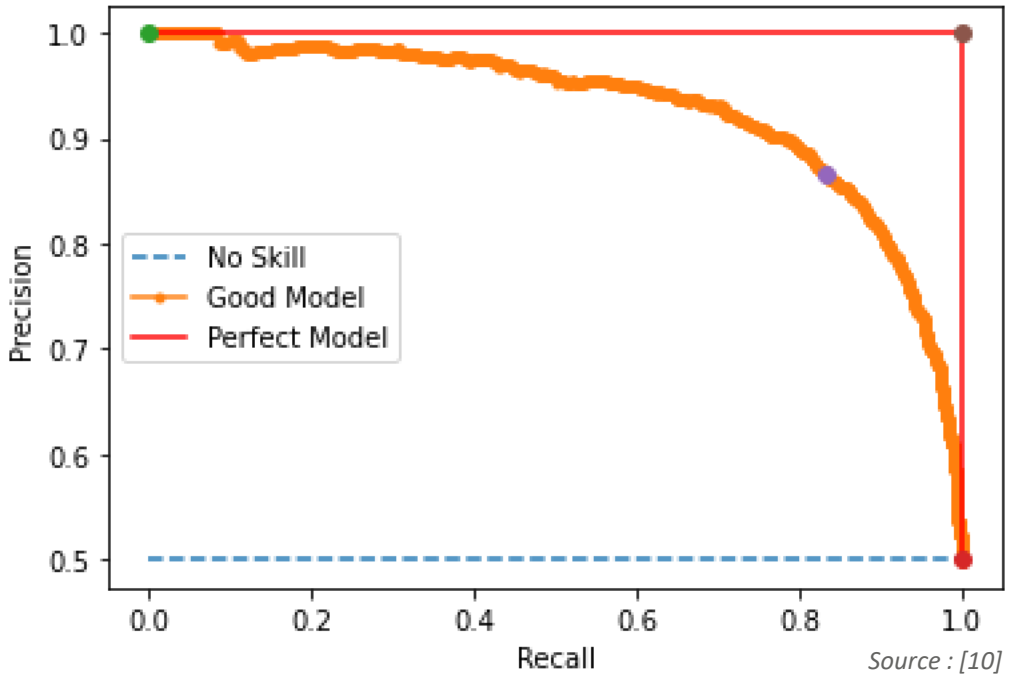


Source : [9]

Métriques d'évaluation – Classification binaire (2/2)

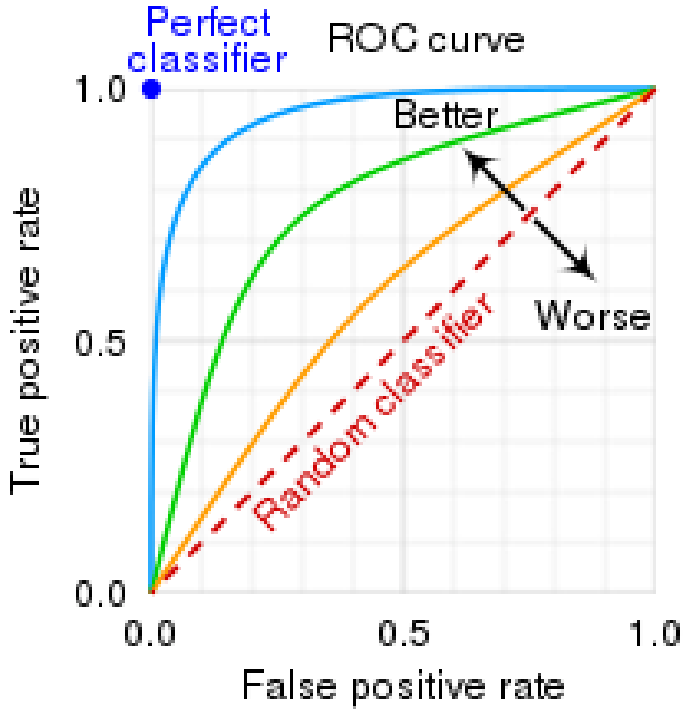
Modèles de classification supervisés :

> Courbe precision recall



Source : [10]

> Courbe ROC

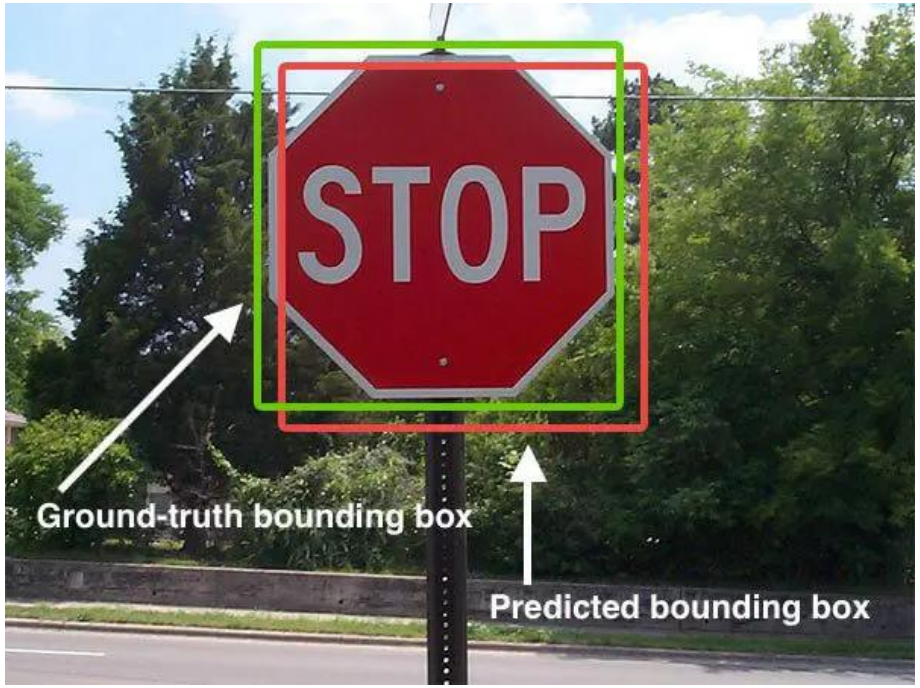


Source : [11]

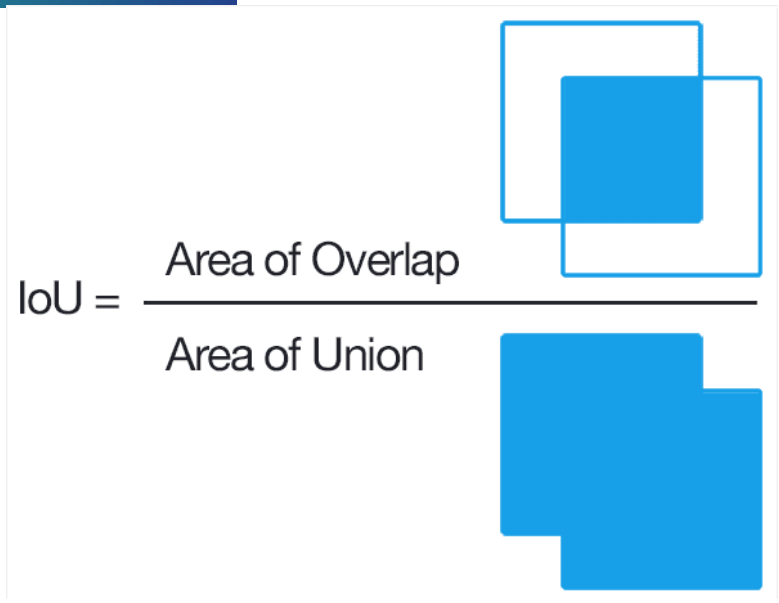
$$TFP = \frac{FP}{FP + VN} \qquad TVP = \frac{VP}{VP+FN} = \text{Rappel}$$

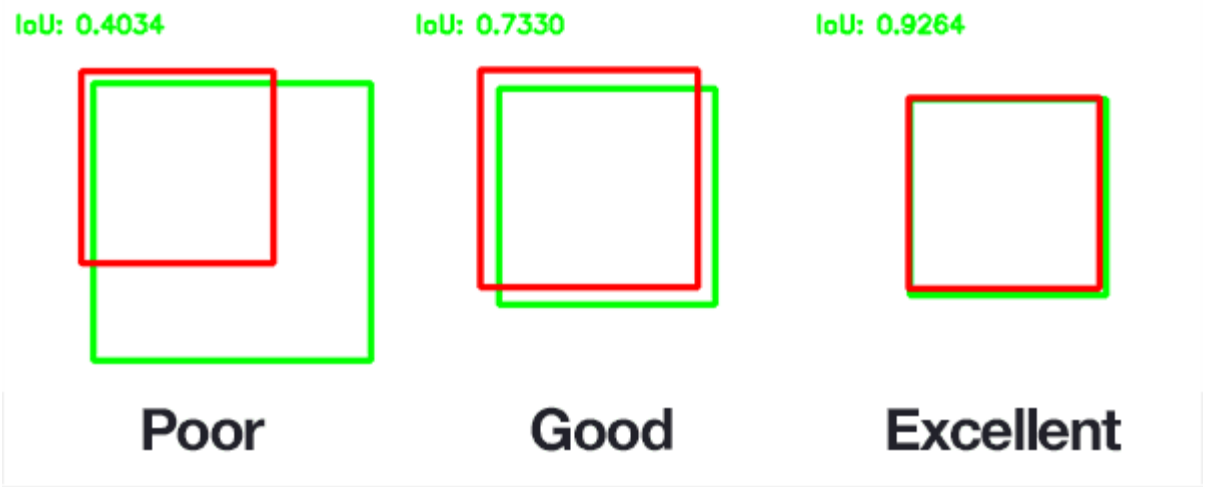
Métriques d'évaluation de l'IA – Détection d'objets

Intersection over Union (IoU)



Source : [12]





Source : [12]

Cas d’usage : **Attribution de crédit** (classification binaire)

- Préparation des données
- Entrainement d’un réseau de neurones
- Evaluation du modèle

Source : [13]

Live coding : Données, entrainement et évaluation

- Importance & défis du test de l’IA
- Propriétés des modèles de ML
  - Ethique & équité
  - Explicabilité
  - Robustesse

Test de l’IA

Pourquoi tester l'IA ?

L'évaluation des performances n'est pas suffisante…

L'évaluation du modèle…

- vérifie qu'il **généralise bien** (pas de surajustement ou de sous-ajustement),
- assure que la **performance globale est satisfaisante**

mais…

- ne localise pas et ne caractérise pas les **erreurs**,
- ne traque pas les **régressions comportementales**,
- ne détecte pas les **biais**,
- etc.

Source : [14]

C'est un détecteur de neige…

Les défis du test de systèmes IA

Problématiques génériques au test

- Manque de spécifications
- Problème de l'oracle de test
- Le rôle du testeur

➔ Des connaissances métiers sont souvent requises

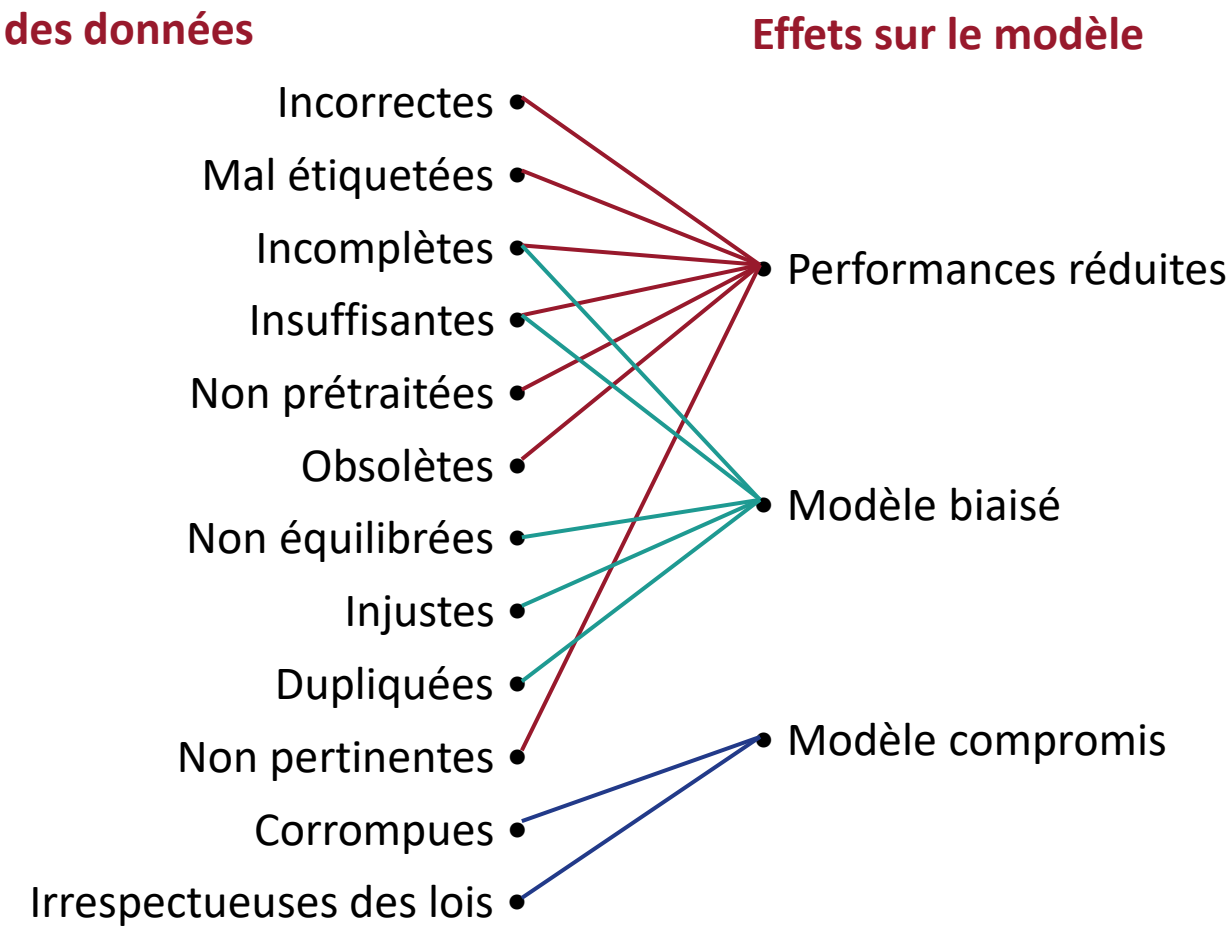
Problématiques spécifiques au ML

- **Métriques d'adéquation des tests** (ex. métriques de couverture)
- Tests de (non) régression
  - Le comportement change à chaque entrainement
- **Problématiques spécifiques au domaine d'application** (ex. adversarial examples, fairness)

IA de confiance

Source : [15]

Qualité des données



Ethique & équité

Principes éthiques

- Respect de l'autonomie humaine
- Prévention de toute atteinte
- Equité
- Explicabilité

→ Accorder une attention particulière aux situations concernant des groupes plus vulnérables

→ Être conscient des avantages et risques liés aux systèmes IA

Source : [15]

Ethique & équité

Biais liés aux données, algorithmes, utilisateurs

- De répercussion ;
- De sélection ;
- D'attribution de groupe ;
- Implicite ;
- D'automatisation/complaisance.

Méthodes de résolution

- Pre-processing :
  - Ajuster les distributions dans les données d'entrainement.
- In-processing :
  - Prendre en compte le critère d'équité dans l'optimisation du modèle.
- Post-processing :
  - Calibrer les prédictions après la phase d'entrainement.

Fairness Goal	Training Scheme	Sensitive Attribute	Method
Disparate impact	Regularization	Known	Prejudice index [45], Absolute correlation [12] Wasserstein fair [41], Pairwise comparisons [11] Fair regression [3], Fair decision trees [4]
	Constraint optimization	Known	Flexible mechanism [83], Tractable constraints [82], Convex-concave [73]
		Unknown	DRO [38], ARL [53], Proxy Fairness [37]
		Beyond	Paper matching [50]
Disparate treatment	Regularization	Known	Controlled direct effect [25], Fair decision trees [4], Convex fair regression [9]
	Constraint optimization	Known	Fairness Through Awareness [29], Logit pairing [35] Counterfactual fairness [52]
Hybrid methods	Constraint optimization	Known	subgroup fairness [46], rich subgroup fairness[46] Maxmin-Fair Ranking [34]

Goal	Training Techniques	Method
Impact-Driven	Adversarial Learning (Prediction-layer)	Adversarial Debiasing [84] Adversarial Recidivism Application [77]
	Adversarial Learning (Hidden-layer)	Censored Representation [30] Transferable Representations [58] Re-embeds word vector [75], Fair Word Embedding [31]
	Disentanglement (Hidden-layer)	Flexibly Fair [36] Disentangled Representations [56]
Treatment-Driven	Contrastive Learning	Contrastive Debiasing [19], Multi-CLRec [86] Conditional Contrastive [76], Fair Graph [51] Fairness-aware Data Augmentation [51]
Hybrid methods	Disentanglement	Flexibly Fair [36], Intersected Fair [63]
	Adversarial Learning	Fair Graph Embeddings [15]

Source : [16]

Explicabilité de l'IA (XAI)

La facilité avec laquelle les utilisateurs peuvent déterminer pourquoi ou comment le système basé sur l'IA produit un résultat particulier.

Objectifs

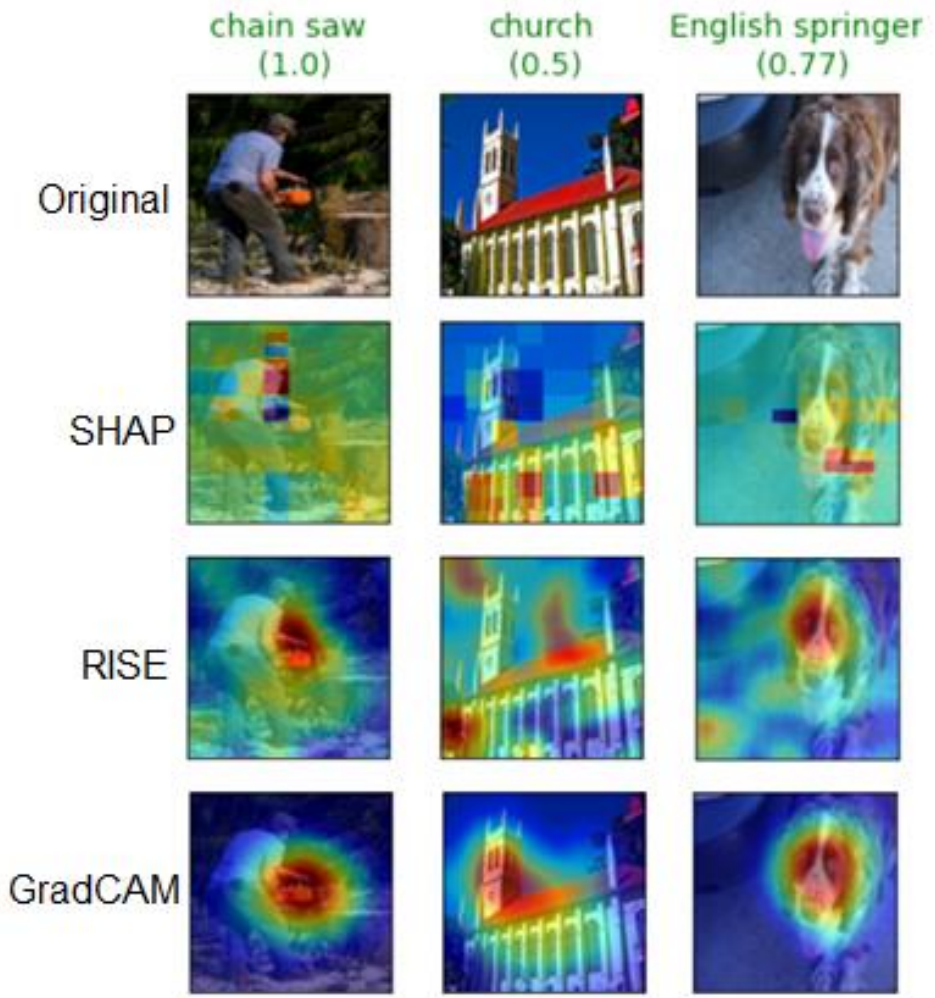
- Justifier pour instaurer la confiance des différents acteurs
- Contrôler pour respecter les normes et exigences, identifier les biais, risques, vulnérabilités
- Améliorer la conception grâce à une meilleure compréhension

Caractéristiques des méthodes

- Modèle : spécifique, agnostique
- Portée des explications : locale, globale
- Types de données : images, texte …
- Technique : simplification, perturbation …

Source : [17]

Explicabilité de l'IA – Classification d'images



Robustesse

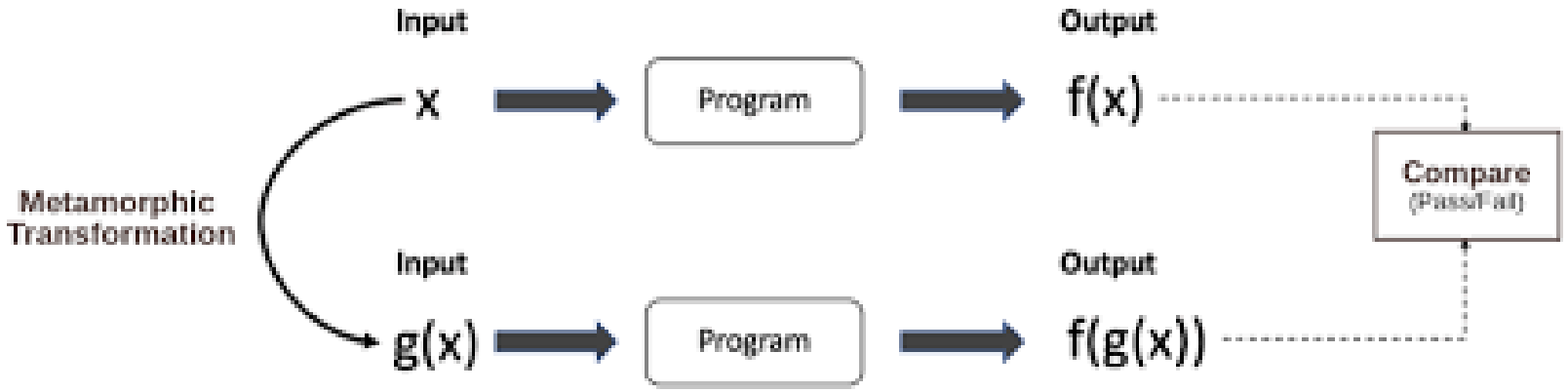
"Ability of an AI system to **maintain its level of performance under any circumstances**."

ISO/IEC TR 24029-1:2021

- Englobe à la fois la **sécurité** et la **sûreté**.
  - Sécurité : "The degree to which a component or system protects its data and resources against unauthorized access or use and secures unobstructed access and use for its legitimate users." (ISTQB)
  - Sûreté : "Safety is considered to be the expectancy that an AI-based system will not cause harm to people, property or the environment." (ISTQB)

- A nuancer : Les circonstances doivent être comprises dans le domaine opérationnel.

Test métamorphique



Source : [19]

Source : [19]

Adversarial testing

Attaque consistant à “**perturber subtilement les entrées valides** qui sont transmises au modèle formé pour l'amener **à fournir des prédictions incorrectes**”. (ISTQB)

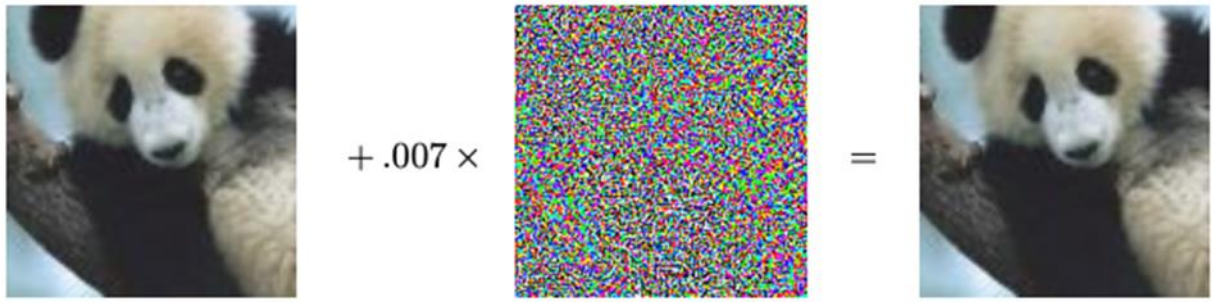
Caractéristiques d’une attaque

- Attaque de disponibilité, d'intégrité ou contre la confidentialité et la vie privée
- Attaque ciblée ou non ciblée
- Attaque causale ou exploratoire
- Attaque boîte blanche ou boite noire

Métriques d’évaluation et d’analyse

- Taux de réussite
- Accuracy du modèle attaqué
- Distortion
- Nombre d’itérations

Fast Gradient Sign Method (FGSM)

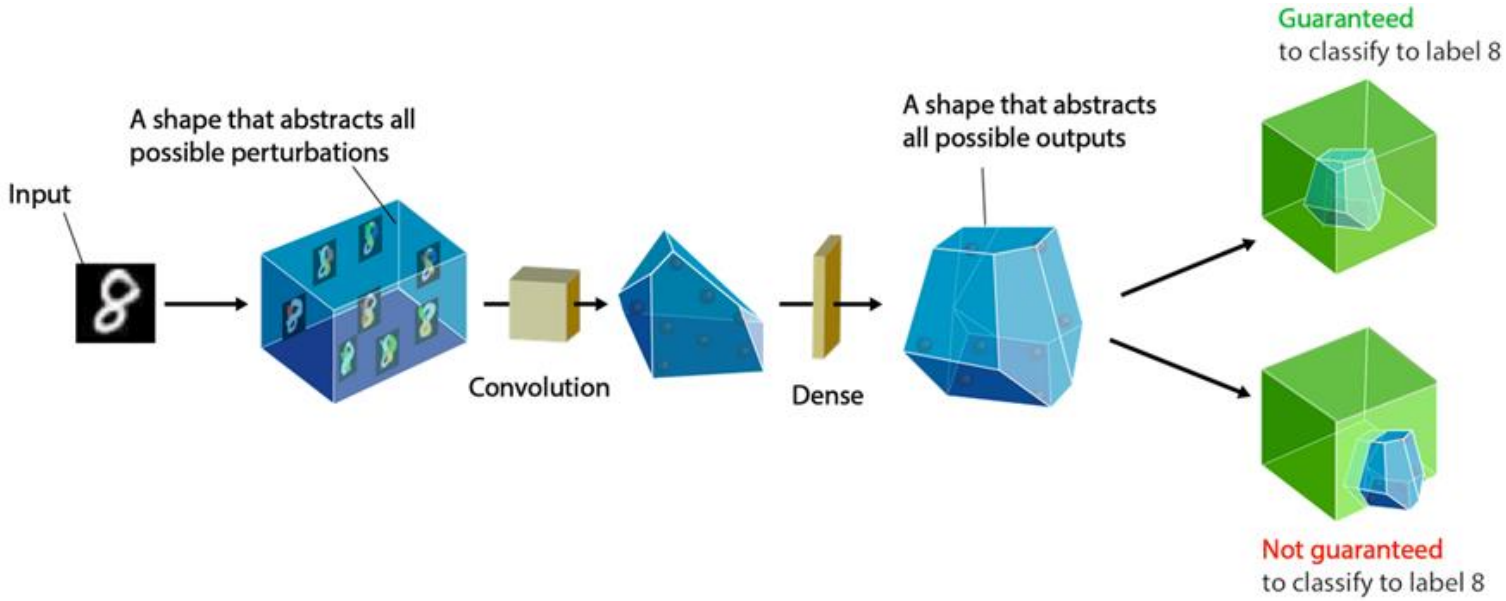


Source : [20]

Autres méthodes

Exemples : méthodes de test et de génération de cas de test

- Vérification formelle
- Test de mutation
- Fuzzing
- Approche « search-based »
- Générateur basé sur un GAN
- Générateur manuel
- …



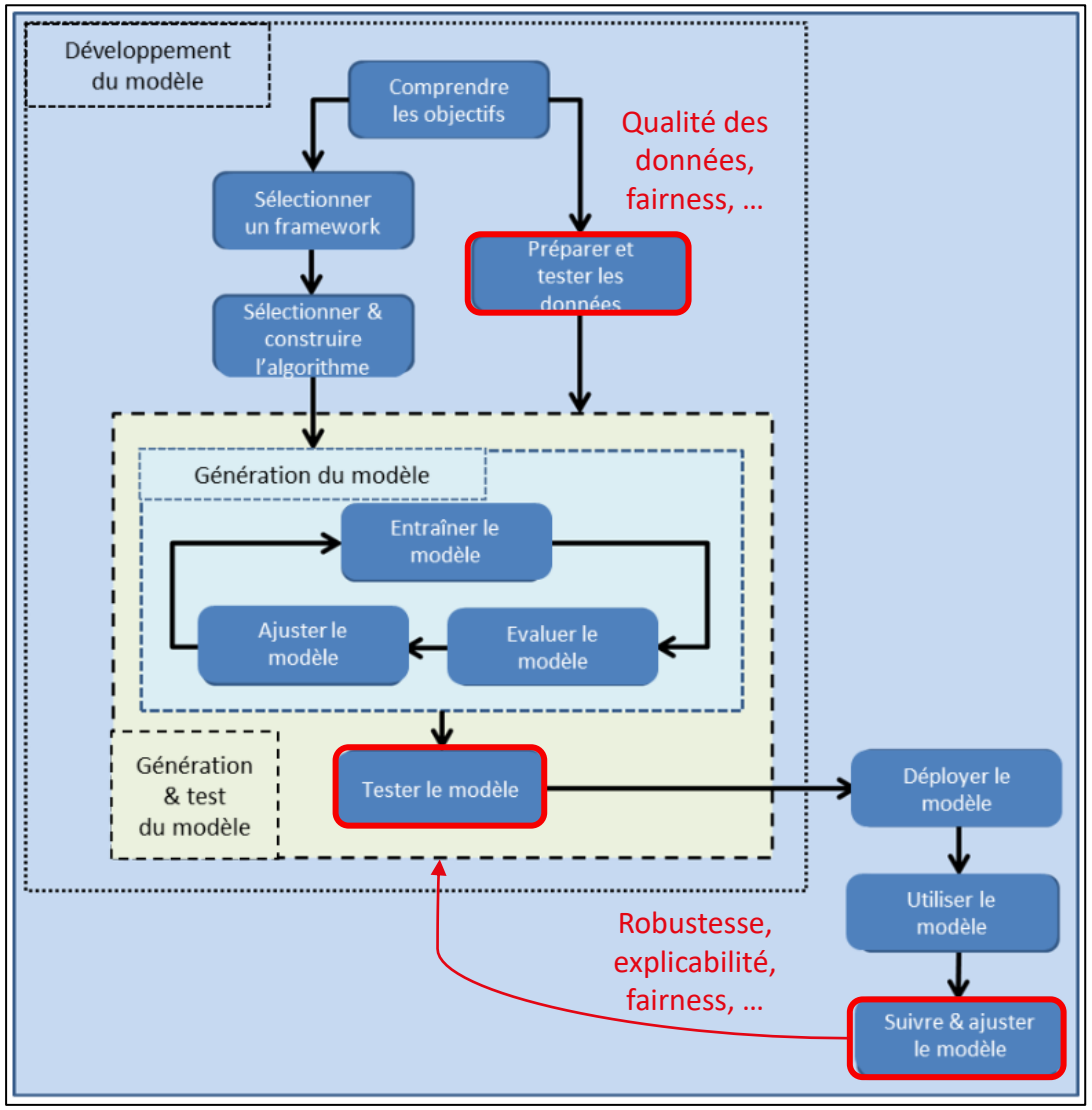
Source : [21]

- Robustesse
- Explicabilité
- Ethique

Live coding : Test

Conclusion

- Connaissances et bonnes pratiques
  - Maîtriser le processus de développement d'un système IA
  - Comprendre les spécificités du test de l'IA par rapport au test logiciel
  - Connaitre les principales faiblesses d'une IA
  - Identifier les méthodes et outils de test pertinents
  - Faire auditer son système IA (AI Act)



Source : ISTQB CT-AI Syllabus v1.0 [7]

Kereval

Laboratoire d’ingénierie du test logiciel

https://www.kereval.com/

contact@kereval.com

sarah.leroy@kereval.com

clement.francois@kereval.com

Sources

- [1] https://www.europarl.europa.eu/news/fr/headlines/society/20200827STO85804/intelligence-artificielle-definition-et-utilisation
- [2] https://openai.com/blog/chatgpt
- [3] https://www.marktechpost.com/2022/11/14/how-do-dall%C2%B7e-2-stable-diffusion-and-midjourney-work/
- [4] https://www.bercynumerique.finances.gouv.fr/quest-ce-que-lia
- [5] https://blent.ai/blog/a/apprentissage-supervise-definition
- [6] https://teahouse.fifty-five.com/en/machine-learning-reinforcement-learning/
- [7] https://www.cftl.fr/wp-content/uploads/2022/08/ISTQB_CT-AI_Syllabus_v1.0-FR.pdf
- [8] https://bigwater.consulting/2019/04/08/software-development-life-cycle-sdlc/
- [9] https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021
- [10] https://analyticsindiamag.com/complete-guide-to-understanding-precision-and-recall-curves/
- [11] https://en.wikipedia.org/wiki/Receiver_operating_characteristic
- [12] https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/
- [13] https://www.fibe.in/blogs/fico-score-vs-credit-score-what-is-the-difference/
- [14] https://medium.com/trusted-ai/explaining-ai-model-behaviour-with-ibm-watson-openscale-86515702c177
- [15] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- [16] https://arxiv.org/abs/2111.03015
- [17] https://ieeexplore.ieee.org/document/8466590
- [18] https://dl.acm.org/doi/10.1145/3340482.3342741
- [19] https://dl.acm.org/doi/abs/10.1145/3241979
- [20] https://arxiv.org/abs/1412.6572
- [21] https://github.com/eth-sri/eran