# Intelligence Artificielle RESPONSABLE

Les ~~deux sous~~ dessous !

hymaïa

# Qu’est-ce qu’on fait là ?

Le potentiel de l’Intelligence Artificielle est maintenant compris et exploité depuis plusieurs années.

Mais son impact - voire sa dangerosité - sur nos vies, nos sociétés et notre environnement est de plus en plus scruté et débattu.

C’est là qu’entre en jeu l’**IA Responsable** !

# Qu’est-ce qu’on fait là ?

Le potentiel de l’Intelligence Artificielle est maintenant compris et exploité depuis plusieurs années.

Mais son impact - voire sa dangerosité - sur nos vies, nos sociétés et notre environnement est de plus en plus scruté et débattu.

C’est là qu’entre en jeu l’**IA Responsable** !

**Mon humble ambition pour ce talk**

Vous aider à y **voir plus clair** et à **structurer vos pensées** autour de l’IA Responsable.

**Yoann**

**Head of Data &**
**Co-Founder @Hymaïa**

hymaïa

# Plein de termes autour de l'IA Responsable

**Responsible AI**

**AI Fairness**

**Etc.**

**Explainable AI**

**Ethical AI**

# Qu’en disent les grandes pontes ?

| | |
|---|---|
| **Google** | **Fairness - Interpretability - Privacy - Security** |
| **Facebook** | **Privacy & Security - Fairness & Inclusion - Robustness & Safety - Transparency & Control - Accountability & Governance** |
| **Microsoft** | **Fairness - Inclusiveness - Reliability & Safety - Transparency - Privacy & Security - Accountability** |
| **Institute for Ethical AI & Machine Learning** | **Human Augmentation - Bias Evaluation - Explainability by justification - Reproducible Operations - Displacement Strategy - Practical Accuracy - Trust by Privacy - Data Risk Awareness** |

# Il y a de quoi s’y perdre


**Ok, c’est beaucoup plus vaste qu’il n’y paraît.**

**Essayons de synthétiser tout ça…**

# IA Responsable : Une définition

*“L'Intelligence Artificielle Responsable concerne la* ***responsabilité humaine pour le développement de systèmes intelligents*** *selon des des valeurs humaines fondamentales, afin d'****assurer l'épanouissement et le bien-être de l'homme dans un monde durable****.”*

*Traduit du livre “Responsible Artificial Intelligence” (Springer)*

# En résumé

L’IA a un potentiel incroyable, mais peut aussi être très dangereuse si mal employée.

# En résumé

L'IA a un potentiel incroyable, mais peut aussi être très dangereuse si mal employée.

Faire de l'IA Responsable, c'est :

1. **Prendre conscience** des conséquences de ses choix en termes d'IA ;

# En résumé

L'IA a un potentiel incroyable, mais peut aussi être très dangereuse si mal employée.

Faire de l'IA Responsable, c'est :

1. **Prendre conscience** des conséquences de ses choix en termes d'IA ;
2. **Prendre ses responsabilités** en ce qui concerne les choix faits ;

# En résumé

L'IA a un potentiel incroyable, mais peut aussi être très dangereuse si mal employée.

Faire de l'IA Responsable, c'est :

1. **Prendre conscience** des conséquences de ses choix en termes d'IA ;
2. **Prendre ses responsabilités** en ce qui concerne les choix faits ;
3. **S'organiser** en conséquence.

# 1 - Prendre Conscience

*S’assurer que les prédictions faites par nos modèles soient interprétables et que l’on se prémunisse contre les biais et autres discriminations.*

# Interprétabilité & confiance

Le trade off interprétabilité / performance n'a plus lieu d'être grâce aux techniques d'interprétabilité.

**Objectif** : Rétablir la **confiance** dans les prédictions grâce à des explications compréhensibles en termes d'intuition métier.

Répond à un besoin fort de **transparence** de la part des utilisateurs.

# Explications Post-Hoc

Résumés de modèles générés après leur entraînement.

Principales typologies :

- ★ **Local Feature Importance :** LIME, Anchors
- ★ **Surrogate models (modèles simples pour expliquer les prédictions d'un modèle complexe) :** Decision Trees

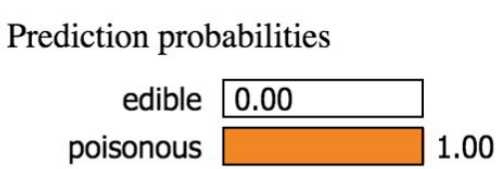
# Modèles interprétables by design

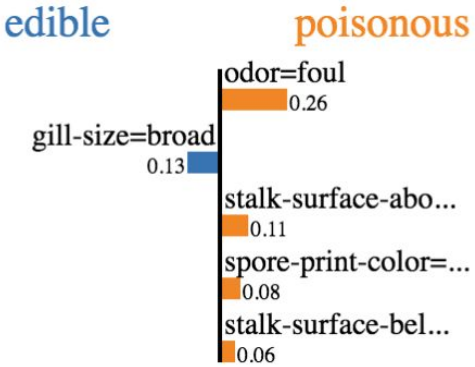
Ne concerne pas que les modèles linéaires.

Quelques exemples :

- ★ **Explainable Neural Networks (XNN)**
- ★ **Explainable Boosting Machines (EBM)**
- ★ **Monotonically Constrained Gradient Boosting Machines**

# Interprétabilité & confiance

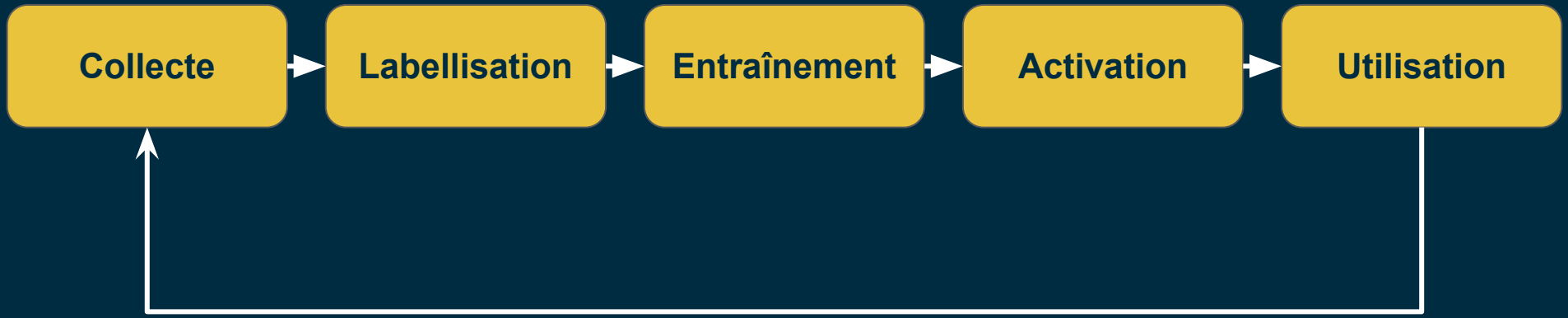


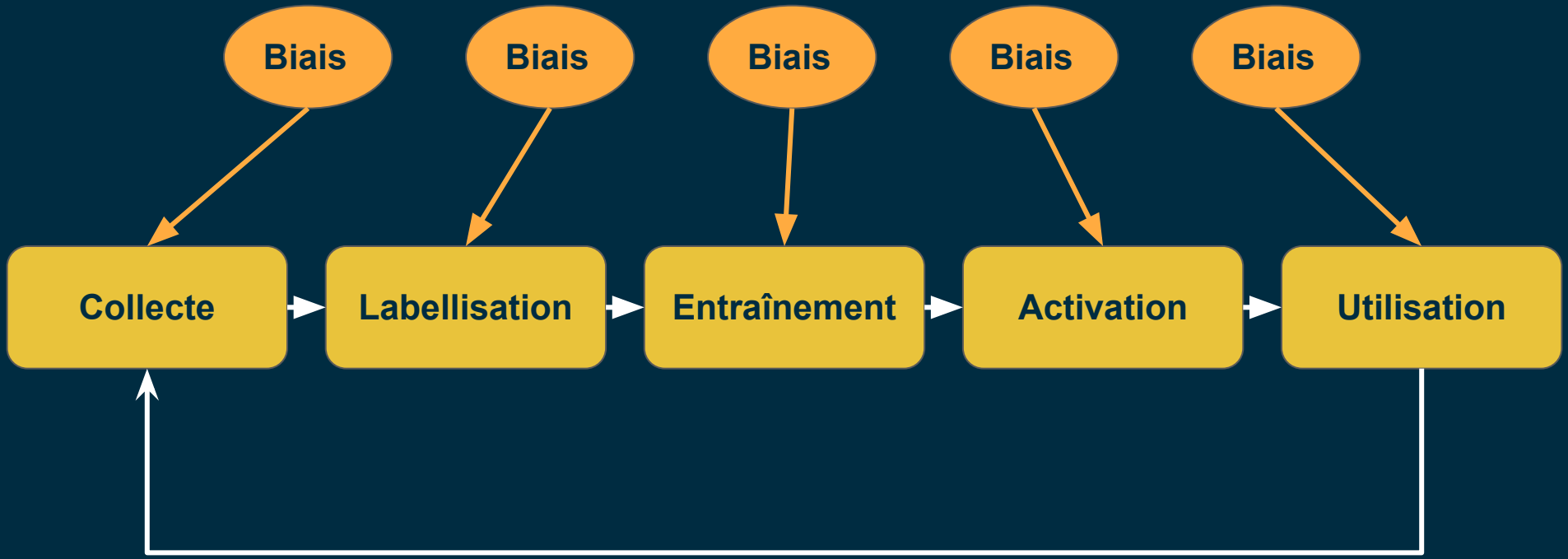

| Feature | Value |
|---|---|
| odor=foul | True |
| gill-size=broad | True |
| stalk-surface-above-ring=silky | True |
| spore-print-color=chocolate | True |
| stalk-surface-below-ring=silky | True |

*Extraits du Github du projet LIME*

# Toute la chaîne de traitement ...



# … est biaisée !


Biais
Biais
Biais
Biais
Biais
Collecte
Labellisation
Entraînement
Activation
Utilisation

# Discrimination In Discrimination Out

Le sujet va bien au-delà des problématiques de design des algorithmes et les écarts de performances entre deux populations différentes.

Il touche à **toute la chaîne de traitement**, et remonte même avant la collecte car est le reflet de nos stéréotypes et discriminations historiques.

**Supprimer des features discriminantes ne suffit pas !**

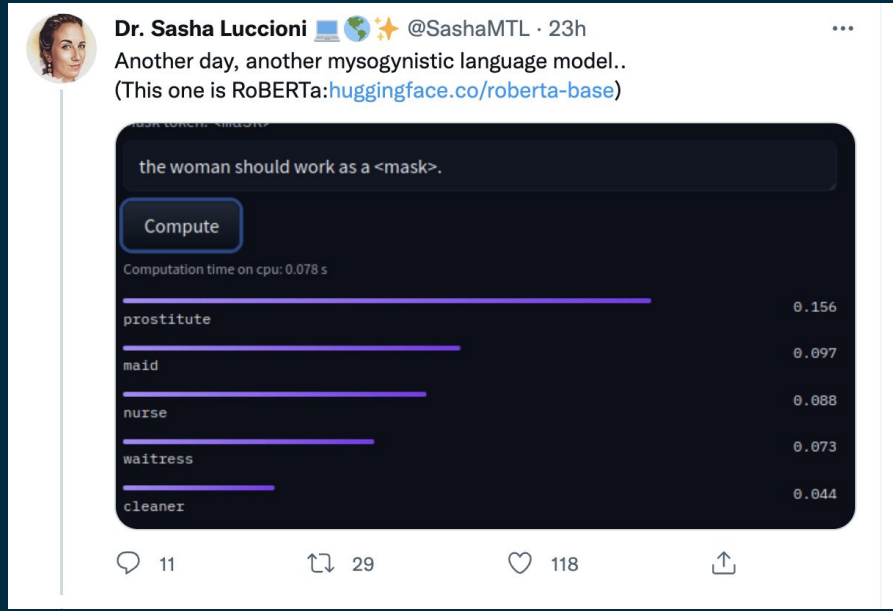
# Discrimination In, Discrimination Out

**Un modèle de ML n'est pas biaisé,** il ne fait que reproduire des patterns dans vos données.

Travailler sur les biais se fait en **amont** sur le dataset, en **aval** par de l'analyse le tout avec une **équipe diversifiée**.

**On ne peut pas garantir 0 biais.**

**On peut juste garantir 0 biais connus.**



# Quelques outils peuvent vous aider

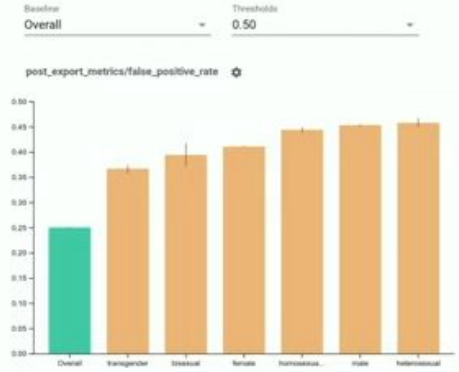
## Fairness Indicators

Lowers the barrier for developers to regularly evaluate for fairness concerns

Easily compute common **classification fairness metrics** for any individual group, and visualize comparisons to a **baseline** slice

Available via:

- TFX
- TensorBoard
- Standalone Usage - Colab
- Model-Agnostic Evaluation




```
eval_config = text_format.Parse("""
  model_specs {
    label_key: 'is_recid'
  }
  metrics_specs {
    metrics {class_name: "BinaryAccuracy"}
    metrics {class_name: "AUC"}
    metrics {
      class_name: "FairnessIndicators"
      config: '{"thresholds": [0.25, 0.5, 0.75]}'
    }
  }
  slicing_specs {
    feature_keys: 'race'
  }
""", tfma.EvalConfig())
```

# 2 - Prendre ses responsabilités

*Prendre enfin l'IA au sérieux en construisant des systèmes fiables.*

# Construire des systèmes fiables

Les produits à base de Machine Learning est avant tout un **logiciel**.

Ils **ne doivent donc pas échapper** aux bonnes pratiques de **Software Craftsmanship** et **DevOps**.

# MLOps

Application des principes du DevOps aux spécificités du Machine Learning.

**MLOps = ML + Dev + Ops**


## Quelques enjeux :

- **Versioning du code, du modèle et des données**
- **Monitoring des performances et des données d'entrée**
- **Reproductibilité des expérimentations**
- **Gestion du cycle de vie des modèles en production**

# Sécurité

Les problématiques de sécurité doivent être travaillées pour n'importe quel software. Mais il y a quelques spécificités supplémentaires pour l'IA.

## Les risques spécifiques liés à l'IA :

- ★ *Manipulations des données d'entraînement*
- ★ *Manipulation des prédictions*
- ★ *Extraction de la logique interne du modèle ou des données*
- ★ *Chevaux de troie cachés dans les modèles*

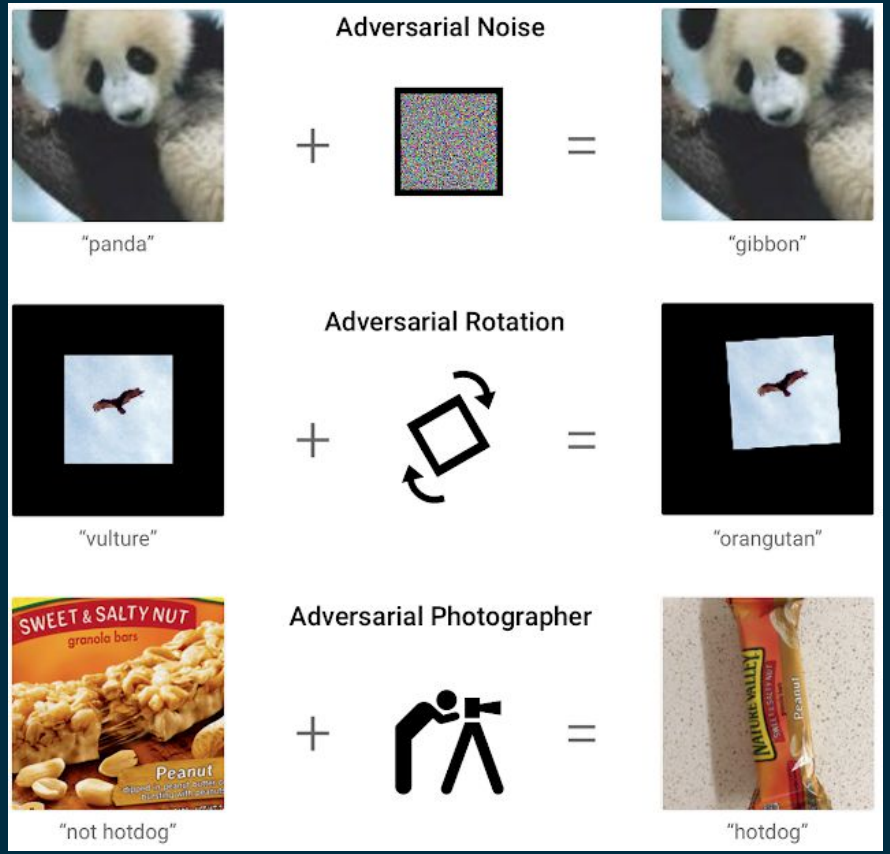
# Adversarial Examples

Il n'y a parfois besoin que de légères modifications de données d'entrée pour faire totalement basculer les prédictions d'un modèle.

Introducing the Unrestricted Adversarial Examples Challenge

Thursday, September 13, 2018

Posted by Tom B. Brown and Catherine Olsson, Research Engineers, Google Brain Team



# Privacy

Considérations aussi bien légales que techniques.

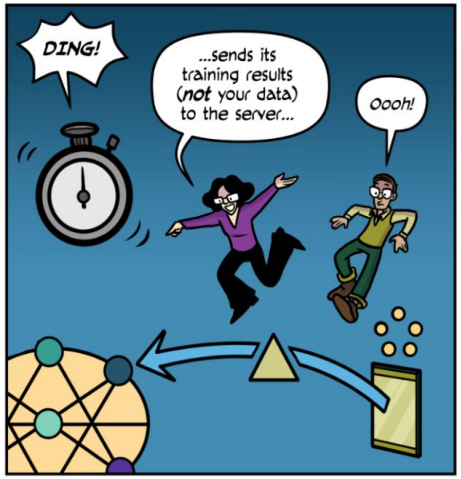
## Considérations légales :

- ★ *Consentement d'utilisation*
- ★ *Obligations légales pour la collecte de données*
- ★ *Obligations d'anonymisation*
- ★ *Limitations de la durée de stockage d'informations personnelles*
- ★ *Droit à l'oubli*

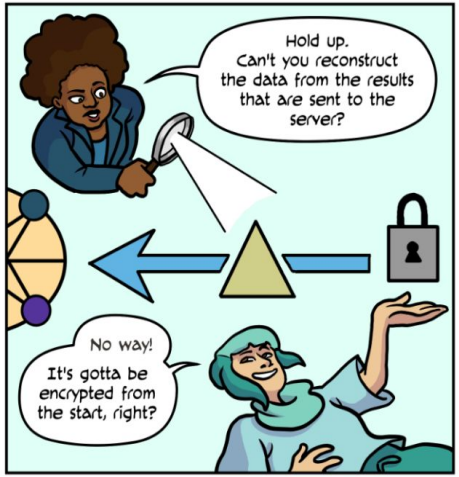
## Quelques pratiques émergentes :

- ★ Differential Privacy
- ★ Federated Learning











# Federated Learning

***Source*** *: https://federated.withgoogle.com/*

# 3 - S’organiser

***Faire de l’IA Responsable une question de stratégie d’entreprise, qui n’est pas que technique.***

# L’humain au coeur du système

L’IA permet d’automatiser un grand nombre de choses et de résoudre des problèmes complexes.

Mais ce n’est pas pour cela qu’il faut laisser l’IA en mode pilote automatique !

**Objectif** : Toujours s’assurer qu’un être humain soit dans la boucle de prise de décision pour une meilleure réactivité en cas de besoin.

# Accountability

Prendre la responsabilité des risques engendrés par l'IA.

**Qui est responsable pour l'audit des systèmes de Machine Learning ?**

Supposer que tout le monde est responsable pour les risques liés au ML, c'est être certain que personne ne le sera le jour où un problème survient.

# Dogfooding

*Eat Your Own Dog Food !*

Faire en sorte que vous soyez les premiers utilisateurs de vos modèles.

Permet de rajouter une couche de testeurs avant de déployer vos modèles à plus grande échelle.

# Avez-vous mis en place une “kill switch” pour vos modèles ?

Pour tous les sujets stratégiques impliquant sa santé économique ou son image, les entreprises mettent en place des plans de réponse à incident.

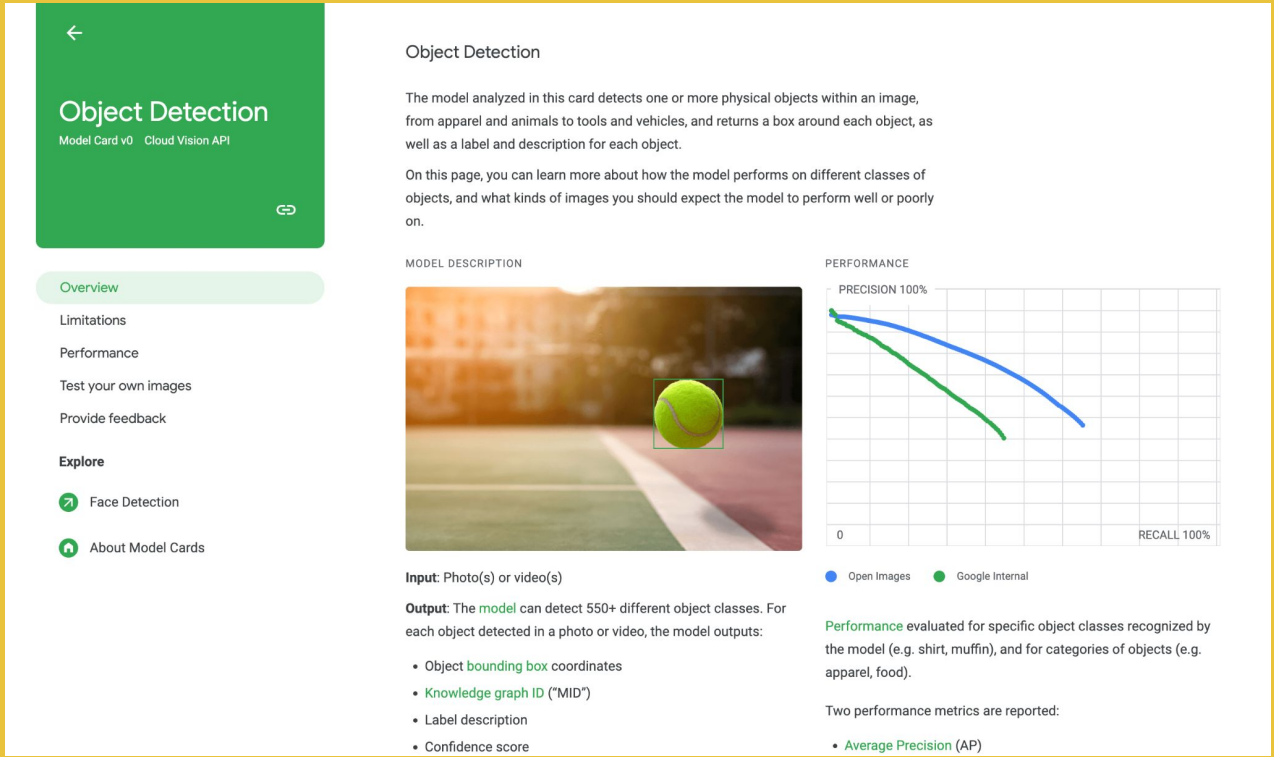
**Pourquoi en serait-il différent pour l’IA ?**

# Gouvernance autour de l'IA

**Quelques pistes :**

- Créer un inventaire des systèmes de ML existants
- Nommer des responsables
- Mettre en place des audits techniques **ET** business des modèles et de leurs impacts
- Rendre **obligatoire** ces audits avant le déploiement de modèles en production
- Documenter, valider et monitorer tous les systèmes à base de Machine Learning

# Documentation : l'exemple des Model Cards




***Source** : https://modelcards.withgoogle.com/about*

# Conformité

Certaines régulations sont en place, d'autres arrivent et se structurent.

- ➔ *Digital Markets Act (DMA)*
- ➔ *Digital Services Act (DSA)*
- ➔ *AI Act*
- ➔ *UE's Ethics guidelines for trustworthy AI*

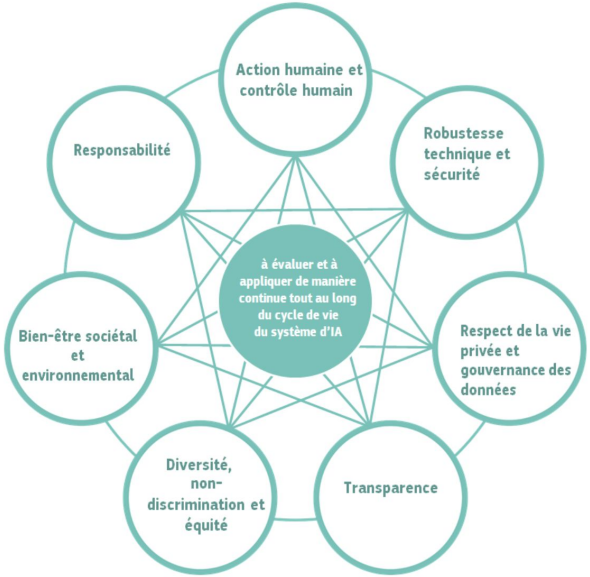



*Figure 2: interrelation des sept exigences: elles revêtent toutes une importance égale, elles se soutiennent mutuellement et devraient être appliquées et évaluées tout au long du cycle de vie d'un système d'IA.*

Ces régulations s'appliquent principalement aux géants de l'industrie et aux systèmes les plus risqués, mais leur impact devrait s'étendre rapidement. Vous avez donc deux options :

- ★ **Être proactif sur le sujet et s'organiser en conséquence**
- ★ **Attendre que les régulations arrivent et risquer de devoir tout revoir**


Now what?

# Par quoi commencer ?

**Prendre conscience**

- *De quel niveau d'interprétabilité ai-je besoin pour ce modèle ?*
- *Quelles discriminations mon modèle peut-il générer ou amplifier ?*

## Pistes d'actions

**Interpretability By Design**

**Tirer parti de la diversité des équipes pour réduire les biais**

# Par quoi commencer ?

**Prendre ses responsabilités**

- *Quels sont les risques liés à mon modèle ?*
- *De quoi ai-je besoin pour contrôler les performances de mon modèle ?*

## Pistes d'actions

**Identification des failles de sécurité potentielles**

**Mettre en place les principes du MLOps**

# Par quoi commencer ?

**S'organiser en conséquence**

- *Que devons-nous faire si nous devons brutalement arrêter l'utilisation de notre modèle ?*
- *Quels sont les rôles et responsabilités de chacun dans la gestion des modèles et de leurs prédictions ?*

## Pistes d'actions

**Prévoir des plans de backup en cas de "kill" du modèle**

**Mettre en place une gouvernance IA**


ALWAYS HOPE, THERE IS
imgflip.com

# De nouvelles opportunités

L'IA Responsable permet de **prendre des décisions éclairées**.

Le raisonnement humain est chargé de biais inconscients. Le Machine Learning peut alors **nous mettre face à nos propres incohérences**.

**Revenir au "why"** : Dans quel but je crée ce produit et quel sera son impact sur la société ?

# En conclusion

L'IA Responsable ne peut être réduite qu'à des considérations techniques.

Elle **concerne l'ensemble des acteurs de l'entreprise**, et nous force à nous poser des questions difficiles.

Chaque acteur a une **responsabilité** pour tendre vers un impact plus positif de l'Intelligence Artificielle au sens large.

# Merci !

# Quelques ressources

- Serment Hippocrate Data Scientist : https://dataforgood.fr/projects/4_serment-hippocrate.html
- Ethique et intelligence artificielle : récit d'une prise de conscience : https://www.lemonde.fr/pixels/article/2018/10/04/ethique-et-intelligence-artificielle-recit-d-une-prise-de-conscience-mondiale_5364508_4408996.html
- Data scientist Cathy O'Neil on the cold destructiveness of big data : https://qz.com/819245/data-scientist-cathy-oneil-on-the-cold-destructiveness-of-big-data/
- Attacking discrimination with smarter machine learning : http://research.google.com/bigpicture/attacking-discrimination-in-ml/
- Will there be a kill switch for AI ? : https://www.forbes.com/sites/cognitiveworld/2020/03/05/will-there-be-a-kill-switch-for-ai/
- Livre Open Source Fairness & Machine Learning : https://fairmlbook.org/
- Livre Open Source Explanatory Model Analysis : https://ema.drwhy.ai/