

INTELLIGENCE ARTIFICIELLE RESPONSABLE

Un panorama des enjeux éthiques autour de l'intelligence artificielle

Grégory Bonnet

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

gregory.bonnet@unicaen.fr

<http://www.gregory.bonnet.free.fr/>



There are certain tasks which computers *ought* not be made to do, independant of whether computers *can* be made to do them.

Joseph Weizenbaum
Computer Power and Human Reason
From Judgement to Calculation
W. H. Freeman, 1976.

1. Du besoin d'une intelligence artificielle responsable

2. Cyberéthique

2.1 Principes fondateurs

2.2 Risques du manque d'éthique

2.3 Exemples d'enjeux pour l'intelligence artificielle

3. Éthique computationnelle

3.1 Agents moraux artificiels

3.2 Problématiques du raisonnement moral

4. Conclusion et bibliographie

Conférence de Dartmouth (1955)

Le problème de l'intelligence artificielle consiste à faire en sorte qu'une machine se comporte d'une manière qui serait qualifiée d'intelligente si un être humain agissait de la sorte.

Journal Officiel (09/12/2018)

Champ interdisciplinaire théorique et pratique qui a pour objet la compréhension de mécanismes de la cognition et de la réflexion, et leur imitation par un dispositif matériel et logiciel, à des fins d'assistance ou de substitution à des activités humaines.

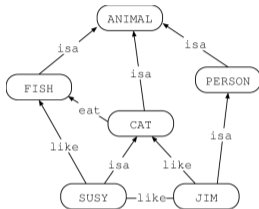
(Floridi, 2023)

1. Reproduire les résultats ou réussites d'un comportement intelligent
2. Produire la source des comportements intelligents par un mécanisme non-biologique

Quatre piliers de l'intelligence artificielle

Un bestiaire de techniques et d'applications

Représenter



Construire des *modèles* qui relient des connaissances entre elles et qui sur lesquels nous pouvons nous appuyer pour prendre des décisions ou résoudre des problèmes.

Résoudre



Concevoir des algorithmes qui, à partir d'un modèle, vont *calculer les meilleures décisions* pour parvenir à un objectif ou une solution à un problème.

Apprendre



Concevoir un modèle à partir d'un *ensemble de données* en s'appuyant sur les corrélations et les similarités entre ces dernières.

Raisonner

Goal reduction phase

$$\frac{}{\Gamma, \Delta \vdash \neg} \neg R \quad \frac{\Gamma, \Delta \vdash B \quad \Gamma, \Delta \vdash C}{\Gamma, \Delta \vdash B \& C} \&R$$
$$\frac{\Gamma, \Delta, B \vdash C}{\Gamma, \Delta \vdash B \Rightarrow C} \Rightarrow R \quad \frac{\Gamma, \Delta \vdash B[y/x]}{\Gamma, \Delta \vdash \forall x.B} \forall R$$

Backchaining phase

$$\frac{\Gamma, \Delta \xrightarrow{D_1} A}{\Gamma, \Delta \xrightarrow{D_1, \&D_2} A} \&L \quad \frac{\Gamma, \Delta \xrightarrow{D_1} A}{\Gamma, \Delta \xrightarrow{D_1, \&D_2} A} \&L \quad \frac{\Gamma, \Delta \xrightarrow{D_1[x]} A}{\Gamma, \Delta \xrightarrow{\forall x, D_1} A} \forall L$$
$$\frac{\Gamma, \Delta_1 \vdash G \quad \Gamma, \Delta_2 \xrightarrow{D} A}{\Gamma, \Delta_1, \Delta_2 \xrightarrow{G \Rightarrow D} A} \Rightarrow L \quad \frac{\Gamma, \vdash G \quad \Gamma, \Delta \xrightarrow{D} A}{\Gamma, \Delta \xrightarrow{G \Rightarrow D} A} \Rightarrow L$$

Identity and Decide rules

$$\frac{\Gamma, D; \Delta \xrightarrow{D} A}{\Gamma, D; \Delta \vdash A} \text{decide!} \quad \frac{\Gamma, \Delta \xrightarrow{D} A}{\Gamma, \Delta, D \vdash A} \text{decide} \quad \frac{}{\Gamma, \vdash A} \text{init}$$

Déduire de *nouvelles connaissances* (preuve à l'appui) à partir d'un modèle en tenant compte des incohérences et des incertitudes.



Il convient de s'assurer que ces systèmes :

- ▶ décident (calculent) et agissent en fonction de facteurs légaux
- ▶ mais aussi culturels, compassionnels et éthiques (subjectifs et pluriels)
- ▶ tout en accomplissant correctement ce pour quoi ils ont été conçus

Cyberéthique

- ▶ Point de vue externe
- ▶ Étude des impacts éthiques de l'intelligence artificielle
- ▶ Régulation
- ▶ Bonnes pratiques

Éthique computationnelle

- ▶ Point de vue interne
- ▶ Développement d'outils d'intelligence artificielle pour l'éthique
- ▶ Implémentation de comportements éthiques

1. Du besoin d'une intelligence artificielle responsable

2. Cyberéthique

2.1 Principes fondateurs

2.2 Risques du manque d'éthique

2.3 Exemples d'enjeux pour l'intelligence artificielle

3. Éthique computationnelle

3.1 Agents moraux artificiels

3.2 Problématiques du raisonnement moral

4. Conclusion et bibliographie

Systèmes d'intelligence artificielle de confiance

Licites, éthiques et robustes

Documents majeurs

- UdeM** Déclaration de Montréal pour une IA responsable (2017)
- CNIL** Comment permettre à l'Homme de garder la main ? (2017)
- CERNA** Éthique de la recherche en robotique et apprentissage machine (2018)
- IEEE** *Global Initiative on Ethics of Autonomous and Intelligent System* (2018)
- UK** *AI in the UK: ready, willing and able?* (2018)
- EU** Lignes directrices en matière d'éthique pour une IA digne de confiance (2019)
- OCDE** Recommandation du conseil de l'OCDE sur l'intelligence artificielle (2019)
- CHN** *The Beijing AI Principles* (2019)
- VAT** *Rome Call for an AI Ethics* (2020)
- UNESCO** Recommandation sur l'éthique de l'intelligence artificielle (2021)

Principes de la bioéthique

- ▶ **Autonomie** : respect du libre arbitre, révocabilité des délégations, innocuité
- ▶ **Bienfaisance** : respect de la dignité humaine, de la vie privée, des environnements et des écosystèmes
- ▶ **Non-malfaisance** : respect de la vie privée, proportionnalité, fiabilité, sécurité
- ▶ **Justice** : équité, non-discrimination, diversité, inclusivité, impartialité

Explicabilité

- ▶ Intelligibilité
- ▶ Transparence
- ▶ Responsabilité

Pourquoi des risques ?

Identifier des principes est insuffisant sans se prémunir des risques de manquements à l'éthique.

Cinq risques fondamentaux

1. **Picorage (Shopping)**. Choisir les principes ou sous-principes qui nous arrangent.
2. **Blanchiment (Bluewashing)**. Se servir des principes éthiques pour ralentir le développement de législations.
3. **Lobbying**. Déclarer sans mise en pratique des principes ou implémenter des mesures superficielles.
4. **Dumping**. Exporter des SIA vers, ou importer des résultats depuis, des pays aux normes moins strictes.
5. **Esquive**. Adapter ses efforts éthiques en fonction du contexte et du rendement perçu.

Différents niveaux de biais

- ▶ Données
- ▶ Algorithmes de décision
- ▶ Mesures d'équité
- ▶ Validation humaine

Problématiques

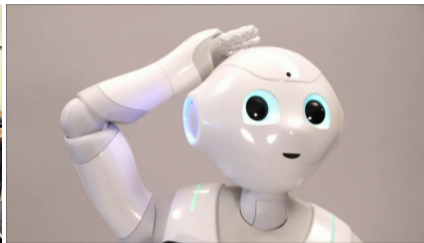
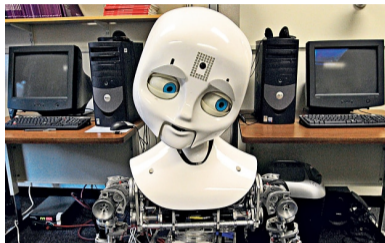
- ▶ **Apophénie.** Corrélations inexactes en raison de biais structurels ou de picorage dans les données
- ▶ **Opacité.** Difficulté à prédire les effets de chaque paramètre (et leur combinaison) sur une décision
- ▶ **Minoration des situations personnelles.** Confrontation de l'individu à une décision fondée sur le collectif
- ▶ **Effets d'ancrage.** Difficulté à réévaluer une décision en apprentissage continu

Coûts énergétiques et environnementaux

- ▶ Production du matériel informatique
- ▶ Stockage des données et des modèles
- ▶ Entraînement (jours et semaines)
- ▶ Requêtes (des millions de requêtes)

Problématiques

- ▶ **Évaluation des impacts.** Quantifier l'impact du matériel et du logiciel (fabrication, stockage et usages)
- ▶ **Usage abusif.** Choix d'une méthode d'intelligence artificielle proportionnée au regard des besoins
- ▶ **Techno-solutionisme.** Développer des systèmes frugaux pour se permettre d'accroître l'usage



Problématiques

- ▶ **Absence de transparence.** Prêter des intentions à une machine en ignorant la nature ou le fonctionnement
- ▶ **Manipulation.** Usage stratégique des émotions pour atteindre un objectif propre à la machine
- ▶ **Coup de pouce (Nudging).** Usage stratégique des émotions pour convaincre un usager humain
- ▶ **Effets émotionnels indirects.** Impacts socio-culturel de l'usage des systèmes d'intelligence artificielle

1. Du besoin d'une intelligence artificielle responsable

2. Cyberéthique

2.1 Principes fondateurs

2.2 Risques du manque d'éthique

2.3 Exemples d'enjeux pour l'intelligence artificielle

3. Éthique computationnelle

3.1 Agents moraux artificiels

3.2 Problématiques du raisonnement moral

4. Conclusion et bibliographie

Objet de l'éthique computationnelle

- ▶ Intégrer des considérations éthiques aux processus décisionnels automatisés
- ▶ Utiliser des outils formels pour représenter et simuler des raisonnements éthiques

Champs de l'éthique en philosophie morale

- ▶ **Méta-éthique.** Statut et sens des concepts éthiques
- ▶ **Éthique appliquée.** Définir et appliquer les règles éthiques d'un environnement particulier (ex. bioéthique)
- ▶ **Éthique normative.** Déterminer, comparer ou expliquer ce qui constitue un comportement éthique

De nombreuses approches prescriptives

- ▶ Approches avec de la théorie des jeux et de la théorie du choix social
- ▶ Approches avec de l'apprentissage automatique et des grands modèles de langage
- ▶ **Approches avec des représentations explicites fondées sur des formalismes**

Agents artificiels moraux

Un AMA est un agent virtuel (logiciel) ou physique (robot) capable soit d'exhiber un comportement moral, soit d'éviter tout comportement immoral. Le terme **moral** est à prendre dans un sens large, fondé sur des théories éthiques, des normes sociales ou tout autre mécanisme de jugement de ce qui est bon ou mauvais.

Construction du modèle interne

- ▶ Agents à modèles descendants, i.e. modèle construit a priori pour s'appuyer sur une théorie morale
- ▶ Agents à modèles ascendants, i.e. modèle par apprentissage
- ▶ Agents à modèles hybrides

Mécanique du modèle interne

- ▶ Agents éthiques implicites, i.e. pas de distinction entre morale et procédures opérationnelles
- ▶ **Agents éthiques explicites**, i.e. représentation explicite interne de la morale
- ▶ Agents pleinement éthiques, i.e. mécanisme de raisonnement méta-éthique (ex. humains)

1. Élicitation morale

Représenter des (systèmes de) valeurs morales et de normes

2. Évaluation morale

Caractériser les conséquences et les responsabilités associées aux décisions

3. Décision morale

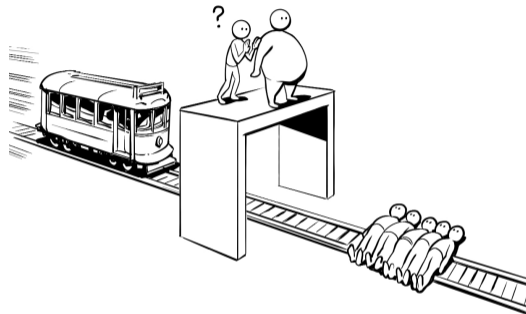
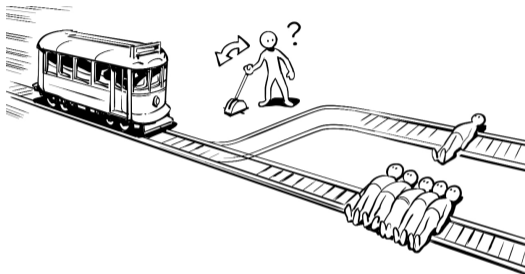
Caractériser la permissibilité ou la nécessité d'une décision donnée.

4. Supervision morale

Juger (vérifier) la conformité éthique ou l'alignement d'un système sur des valeurs morales

Modéliser la responsabilité

Distinguer le dilemme du trolley et du dilemme du pont (cc) David Navarrot



Bien que les dilemmes soient égaux en termes de vie et de mort, ils diffèrent causalement

- ▶ **Trolley** : le mauvais effet est contingent au bon effet
- ▶ **Pont** : le mauvais effet est nécessaire à l'existence du bon effet

1. Du besoin d'une intelligence artificielle responsable

2. Cyberéthique

2.1 Principes fondateurs

2.2 Risques du manque d'éthique

2.3 Exemples d'enjeux pour l'intelligence artificielle

3. Éthique computationnelle

3.1 Agents moraux artificiels

3.2 Problématiques du raisonnement moral

4. Conclusion et bibliographie

Cyberéthique et éthique computationnelle

- ▶ Cinq principes fondateurs qui font consensus
- ▶ Risques de manque éthique (picorage, blanchiment, etc.)

Exemples de problématiques non abordées

1. **Soumission à la machine.** Difficulté à s'opposer à une recommandation ou décision automatisée
2. **Problèmes d'inocuité.** Décisions automatisées ayant un impact irréversible ou difficile à renverser
3. **Vol et pillage de données.** Apprentissage sur des données sans l'accord de leur propriétaire / créateur
4. **Minoration des parties prenantes.** Absence de gouvernance multipartite des systèmes d'IA

Interrogations

- ▶ Comment assurer l'évaluation de l'impact éthique des systèmes ?
- ▶ Comment transcrire dans la loi une politique éthique ambitieuse ?
- ▶ Comment donner des droits aux citoyens vis-à-vis des systèmes d'intelligence artificielle ?

-  **L'intelligence artificielle : de quoi s'agit-il vraiment ?**
Groupe de recherche IA (ouvrage collectif), Cépaduès, 2020.
<https://ia.gdria.fr/>
-  **Comité National Pilote d'éthique du Numérique (CNPEN)**
<https://www.ccne-ethique.fr/fr/cnpen>
-  **Ethics Guidelines for Trustworthy AI**
Commission européenne, 2019
<https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>
-  **Recommandation sur l'éthique de l'intelligence artificielle**
UNESCO, 2022
<https://www.unesco.org/fr/articles/recommandation-sur-lethique-de-lintelligence-artificielle>
-  **Moral Machines**
Wendell Wallach and Colin Allen, Oxford University Press, 2008.
-  **L'éthique de l'intelligence artificielle – Principes, défis et opportunités**
Luciano Floridi, Éditions Mimésis, 2023
-  **Artificial Moral Agents: A Survey of the Current Status**
José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, Félix Ramos
Science and Engineering Ethics, 26:501-532, 2020