Éthique de la recherche en robotique

GdR Robotique – Action Transverse « Robotique et Société »

7 décembre 2022

« Que dois-je faire? »

Raja Chatila

ISIR – Sorbonne Université

La recherche et la technologie

- Effets globaux et à long terme
- Effets cumulatifs
- Effets irréversibles
- Risques existentiels

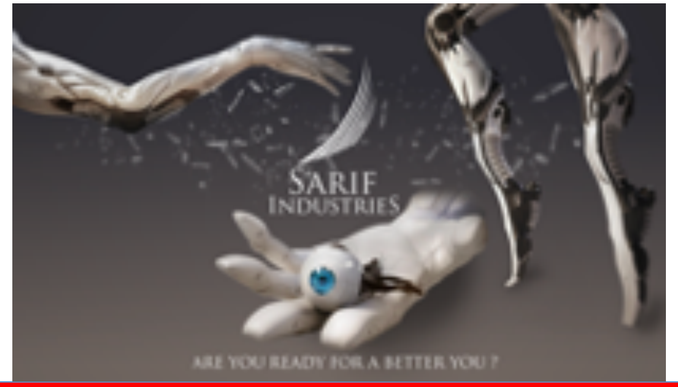
Journal of Personality and Social Psychology
2018, Vol. 114, No. 2, 246–257

0022-3514/18/$12.00 http://dx.doi.org/10.1037/pspa0000098

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



Valeurs

- Les valeurs peuvent être considérées comme des capacités ou des dispositions inhérentes aux choses
- Les valeurs sont des phénomènes clairement perceptibles
- Les valeurs ne sont pas les caractéristiques des choses
- Les choses ne sont que des "porteuses" de valeurs
- Les valeurs sont indépendantes des choses qui les portent;

Max SCHELER, (1874 – 1928)

Der Formalismus
in der Ethik und die Materiale Wertethik,
Elibron Classics. 1921 (2007).

Valeurs intrinsèques

- La valeur intrinsèque est quelque chose a de la valeur "en soi” (Zimmerman 2010).
- La valeur intrinsèque est définie en elle-même (elle ne sert pas à quelque chose).
- Ex: le bonheur est une valeur intrinsèque. «À quoi sert le bonheur?»

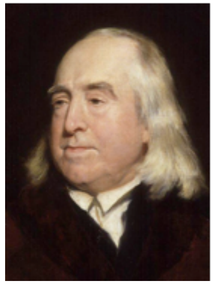
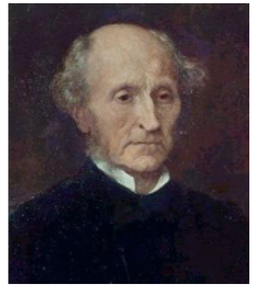
Zimmerman, Michael J. 2010. Intrinsic vs. Extrinsic Value. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab

Valeurs extrinsèques

- Les valeurs extrinsèques ou «instrumentales» mènent causalement à des valeurs intrinsèques.
- Elles ne sont pas bonnes en elles-mêmes, mais se rapportent à, et permettent quelque chose d'autre qui est bon.
- Une valeur extrinsèque tire sa valeur du fait qu’elle conduit au bien (intrinsèque) supérieur (Zimmerman 2010).

Zimmerman, Michael J. 2010. Intrinsic vs. Extrinsic Value. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab

Conséquentialisme/utilitarisme

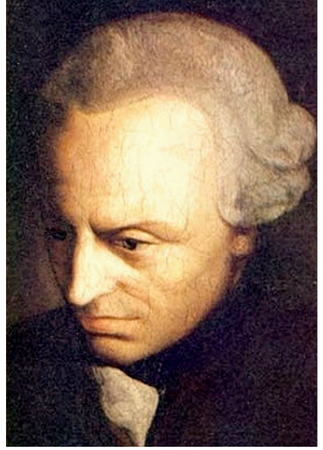



Jeremy Bentham (1789),
John Stuart Mill (1861)

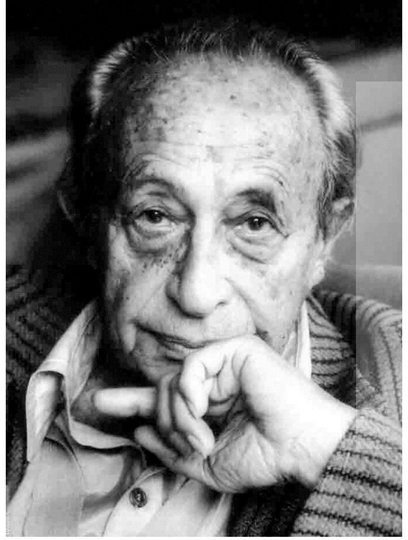
- Utilitarisme: “Le plus grand bien pour le plus grand nombre”

- Seules les conséquences comptent dans un choix moral

Emmanuel Kant
Fondements de la métaphysique des mœurs, 1785



- C’est l’autonomie, la liberté de choix qui rend possible la moralité.
- L’impératif catégorique: la morale est fondée sur une nécessité rationnelle.
- *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.*



Hans Jonas

Le principe de responsabilité [1979]

L’éthique classique

- Toute relation avec le monde non-humain est éthiquement neutre.
- Toute l'éthique traditionnelle est anthropocentrique.
- L'entité "homme" est considérée comme constante dans son essence.
- Faire de bonnes actions ou éviter de faire le mal est associé à l'acte lui-même, ce qui signifie que la portée des actions est limitée dans le temps et l'espace. L'éthique traditionnelle est une éthique du "voisin".

Agis de façon que les effets de ton action soient compatibles avec la permanence d’une vie authentiquement humaine sur terre

Principes d'éthique biomédicale	Principes d'éthique du numérique
• **Principe d'autonomie:** obligation de respecter les capacités de décisions et le consentement des personnes autonomes;	• **Principe d'autonomie**: préserver l'agentivité et le contrôle humains;
• **Principe de bienfaisance:** obligation de procurer des bénéfices et de mesurer les bénéfices par rapport aux risques;	• **Principe de non-malfaisance**: ne pas nuire ni exacerber un mal (sûreté, sécurité, robustesse technique);
• **Principe de non-malfaisance:** obligation d'éviter de nuire;	• **Principe de Justice**: équité, réduction du biais, non discrimination, proportionnalité ;
• **Principe de Justice:** obligation d'équité, juste distribution des bénéfices et des risques.	• **Principe d'explicabilité**: transparence, interprétabilité, traçabilité, auditabilité.

Beauchamp TL, Childress J. Principles of biomedical ethics. New York: Oxford University Press, 1st ed, 1979

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

Exigences pour une IA de confiance

1. **Agentivité humaine et contrôle humain :** droits fondamentaux, action humaine, contrôle humain
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, intégrité des données, accès
4. **Transparence:** traçabilité, explicabilité, communication
5. **Diversité, non-discrimination et équité:** absence de biais injustes, accessibilité, conception universelle, participation des parties prenantes
6. **Bien-être sociétal et environnemental:** durabilité, respect de l'environnement, impact social, démocratie
7. **Responsabilité :** auditabilité, la réduction des incidences négatives et la communication, arbitrages, recours.

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

A risk-based approach to regulation

Proposition AI Act
(21/04/2021)

Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical devices

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

*Not mutually exclusive

AI with specific transparency obligations
'Impersonation' (bots)

Permitted but subject to information/transparency Obligations

Minimal or no risk

Permitted with no restrictions

European Commission

Examples of Potential Harms

Harm to People

- Individual: Harm to a person's civil liberties, rights, or physical safety.
- Group/Community: Harm to a group such as discrimination against a population sub-group.
- Societal: Harm to democratic participation or educational access.

Harm to an Organization/ Enterprise

- Harm to an organization's business operations.
- Harm to an organization from security breaches or monetary loss.
- Harm to an organization's reputation.

Harm to a System/Ecosystem

- Harm to interconnected and interdependent elements and resources.
- Harm to the global financial system, supply chain, or interrelated systems.
- Harm to natural resources, the environment, and planet.

NIST AI Risk Management Framework: 2nd Draft
https://www.nist.gov/itl/ai-risk-management-framework

Méthodologie de conception éthique basée sur les valeurs

Analyse des objectifs, des finalités et des impacts du projet

Choix éclairé de l'approche technique

Découverte des valeurs

Identification des parties prenantes et de leurs valeurs.

Comment sont-elles impactées?

Conceptualisation des valeurs

Quels sont les composantes des valeurs?

Quelles sont les tensions entre elles?

Analyse des valeurs

Quelles sont les priorités entre les valeurs ?

Synthèse de la valeur technique

Solutions techniques pour mettre en oeuvre les choix éthiques et les priorités entre les valeurs