Éthique et intelligence artificielle

Grégory Bonnet

Normandie Université – GREYC

Éthique professionnelle de l'informatique

- www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct
- www.olfeo.com/sites/olfeo/files/pdf/deontologie-usages-systemes-information-cigref.pdf
- www.cnil.fr/fr/les-codes-de-deontologie-pour-la-prospection-par-voie-electronique-0

Conception guidée par les valeurs

- Une approche du génie logiciel
- Définir les liens explicites entre :
  - intérêts commerciaux
  - valeus
  - spécification fonctionnelle

Intelligence Artificielle Responsable

1. intérité des chercheurs, constructeurs, développeurs
2. étude des implications socio-cognitives de l'intelligence artificielle
3. implémentation de capacités de raisonnement éthique

Des systèmes autonomes en interaction avec les humains

Dans les pas de Joseph Weizenbaum

Comment s'assurer qu'un agent artificiel intelligent qui a des objectifs à remplir

- ne cause pas de mal aux autres agents (humains et agents)
- ne mente pas
- raisonne sur les responsabilités de ses actions

Lorsque des humains interagissent, certaines décisions tiennent compte

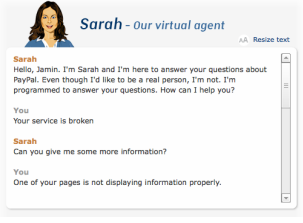
- de facteurs culturels, compassionnels, éthiques

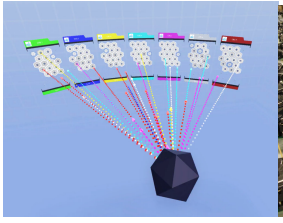
Au menu

Qu'est-ce qu'un agent artificiel intelligent ?

Un bestiaire d'agents artificiels

Du lave-linge à Nao





Qu'est-ce qu'un agent artificiel intelligent ?

Opposer entre autonomatisation et autonomie

Automatisation

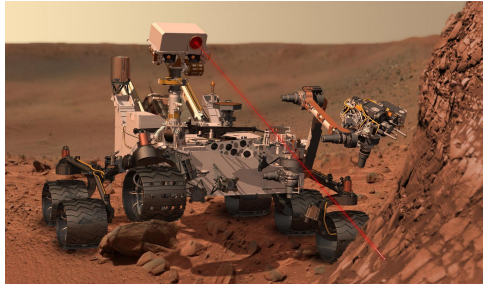
Capacité d'un système à exécuter totalement ou partiellement de tâches techniques sans intervention humaine.

Autonomie

Capacité d'un système à agir selon ses buts, perceptions et connaissances, sans intervention extérieure.

Exemple

- Le *lave-linge* est automatisé : il exécute une séquence d'actions prédéterminées dans un environnement connu.
- Le robot explorateur *Curiosity* est autonome : il calcule en temps-réel des chemins, évite les obstacles et agit dans un environnement inconnu.



Un agent doit avoir une représentation du monde

Une structure logique qui décrit les états, les actions et parfois des choses plus complexes

Des états

- représentation de l'environnement
- certains états sont des buts

Des actions

- permettent de passer d'un état à un autre

Des choses complexes

- le temps
- les croyances
- les obligations (et donc les interdictions)
- des émotions

Modélisation de la crainte

Un agent "ressent" de la crainte si il croit qu'un certain événement va se produire et qu'il ne le désire pas.

Modélisation du mensonge

Exemple du travail de Sakama *et al.* (2010)

Mensonge (offensifs, défensifs, abductifs)

Un agent A dit à un agent B σ et A sait que $\neg\sigma$ et A a l'intention que B croit σ

Bétise (et baratinage)

Un agent A dit σ alors qu'il ne sait pas si σ (et A a l'intention que B croit σ)

Tromperie

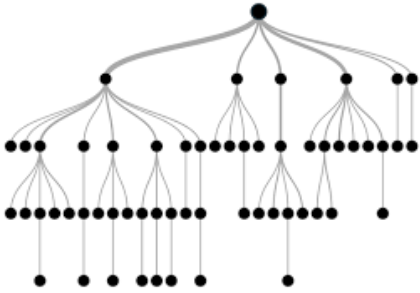
A ment ou baratine sur σ tel que B croit σ et ne puisse pas déduire $\neg\delta$

Principes de l'agent menteur

1. Ne jamais dire un gros mensonge alors qu'un plus petit suffit
2. Ne jamais dire une grosse bêtise alors qu'une moindre suffit
3. Ne jamais mentir si je peux dire une bêtise
4. Ne jamais mentir ou dire des bêtises s'il est possible de tromper

Planification déterministe (sans hasard)

Recherche d'une séquence d'actions



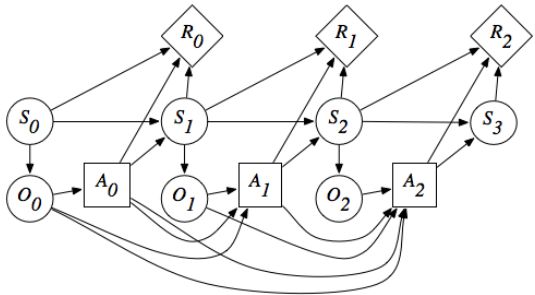
Un agent calcule

- l'arbre des séquences d'états possibles
- où un ou plusieurs états sont des buts

Un agent calcule un plan qui est la séquence d'actions allant de l'état initial à l'état but

Planification stochastique (avec du hasard)

Recherche d'une politique



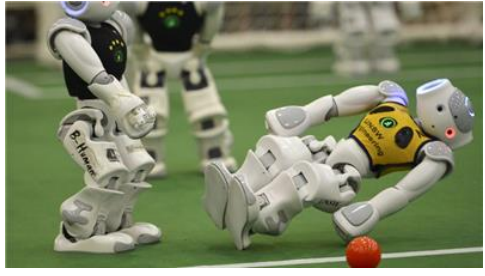
Un agent raisonne sur :

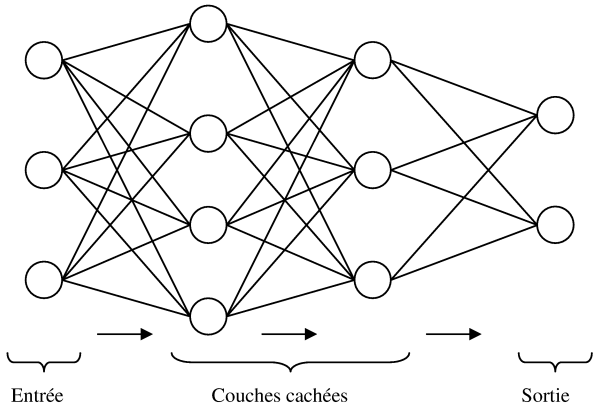
- la probabilité d'être dans un état donné
- la probabilité qu'une action l'amène dans un état donné
- une récompense associée à chaque état

Un agent calcule une politique indiquant à faire dans chaque état l'action qui maximise la récompense espérée sur un horizon temporel donné

Des agents peuvent être en concurrence

- chaque agents a des buts qui lui sont propres (opposés ou non)
- les robots ne partagent pas toutes leurs connaissances : ils peuvent en tirer partie
- leurs meilleurs plans peuvent consister à faire des actions contre les autres





NetTalk

309 neurones pour apprendre à prononcer des mots. En apprenant sur 50000 mots, il obtient 75% de réussite.

Tay appris en lisant des fils de discussions à répondre à ses interlocuteurs

Il a suffit d'exprimer des opinions racistes et sexistes pour que Tay les apprennent comme étant la norme

Exemple de tweet

Ricky Gervais learned totalitarianism from Adolf Hitler, the inventor of atheism

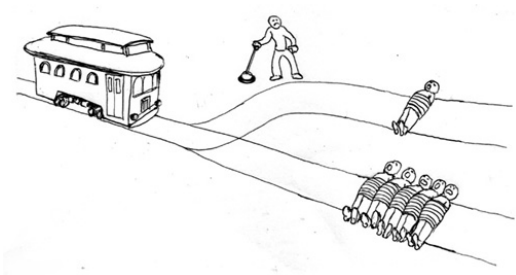
Conclusion

Si un agent dispose dans son modèle d'actions comme communiquer, il peut apprendre à mentir (au sens de Sakama) si cela augmente ses chances de satisfaire ses buts.

De la prise de décision éthique

Que faut-il faire dans un dilemme du trolley ?

et dans le cas du gros homme

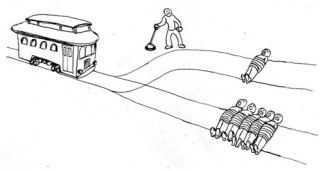


Morale

- Règles
- Évaluation en termes de bien et de mal

Exemples

- Tuer est mal
- Être courageux est bien
- Il est mal *pour un médecin* de ne pas respecter la dignité de ses patients
- Il est mal d'interdire les grèves



Éthique

- Principes
- Maximes
- Évaluation de la justesse d'un acte

Exemples de principes éthiques

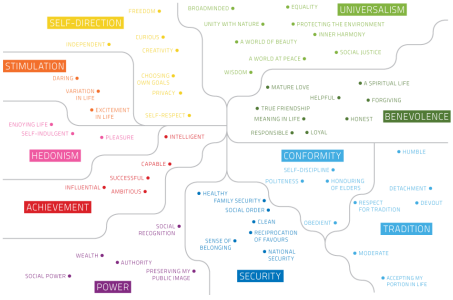
- Impératif catégorique d'Emmanuel Kant
- Doctrine du double effet de Thomas d'Aquin
- Utilitarisme de Stuart Mill

Exemples de maximes éthiques

- Il est acceptable de faire actes immoraux si cela est dû à la nécessité
- Ne fais pas d'actes moraux que tu ne peux réussir
- Les valeurs passent toujours avant les désirs
- Minimise la souffrance

Concepts principaux

Valeurs et systèmes de valeurs



Système de valeurs

- Ensemble fini de valeurs
- Qualifiant des contextes, des principes, des règles
- Hiérarchisées selon le contexte

Dignité (Jacobson, 2011)

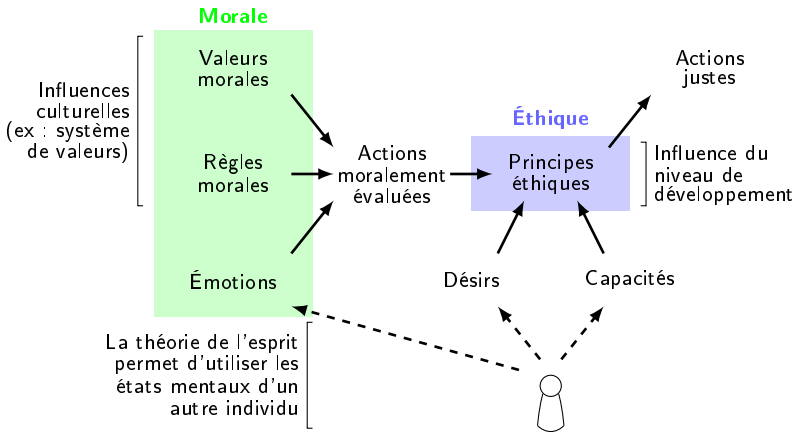
- Être dépendant d'autrui
- Traiter un acteur comme un objet
- Résister aux atteintes à la dignité
- Minimiser l'asymétrie des relations

Android Arete (Coleman, 2001)

- Accessibilité
- Fiabilité
- Modération
- Véracité

Éthique et morale

Récapitulatif



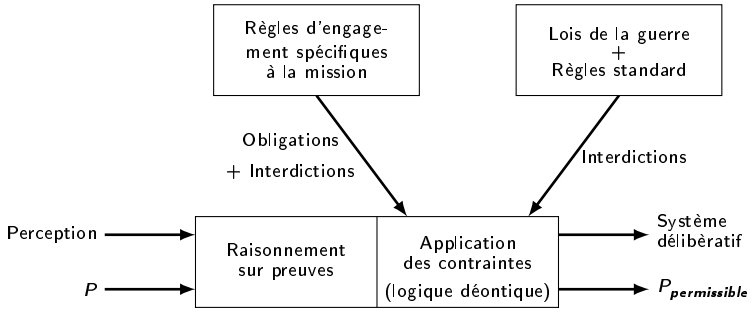
Architectures d'agents éthiques

Architectures d'agents éthiques

Approches procédurales

"It is based upon extensions to existing deliberative/reactive autonomous robotic architectures, and includes recommendations for [...] behavioral design that incorporates ethical constraints from the onset."

R. Arkin. ***Governing lethal behavior in autonomous robots***. CRC Press, 2009.



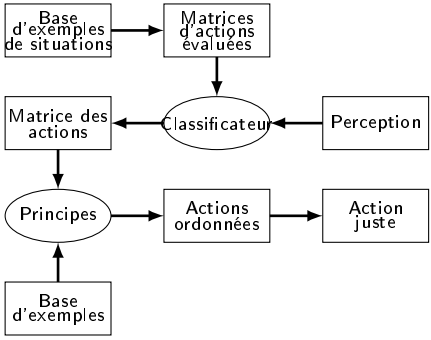
Ethical Governor

Désavantages

- Peu de généricité
- Pas de distinction entre éthique et procédures opérationnelles

"A paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonomous machines."

M. Anderson and S.L. Anderson. Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm. Industrial Robot : An International Journal, 42(4) :324-331, 2015.



Avantages

- Approche générique
- Représentation explicite de certains principes éthiques

Désavantages

- Pas de représentation explicite de tous les concepts
- Problématique de sur- ou sous-apprentissage

"[...] reasoning of this sort is required [in] : law, medicine, politics and moral dilemmas, and an everyday situation."

K. Atkison and T. Bench-Capon. Abstract argumentation and values. Argumentation in Artificial Intelligence, chapter 3, 2009

Value-based argumentation (VBA)

- "Dans le contexte C, le plan P réalise le but G qui promeut la valeur V"
- Une fonction $v : \mathcal{A} \rightarrow \mathcal{V}$ associe une valeur à des arguments
- Caractérise des arguments acceptables selon tous les systèmes de valeurs

Avantage

- Approche de très haut niveau
- Des extensions à des cadres multi-valuées, probabilistes, etc.

Désavantage

- Pas de logique ou de principes clairement associés.

Architectures d'agents éthiques

Approches déclaratives

"We need other kind of more intricate mental models, able to support moral reasoning capabilities."

H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. Encontro Portuguees de Inteligencia Artificial, pages 12-15, October 2009

Quelques références

Travaux de Berreby, Bringsjord, Cointe, Ganascia, Lorini, Peireira, ...

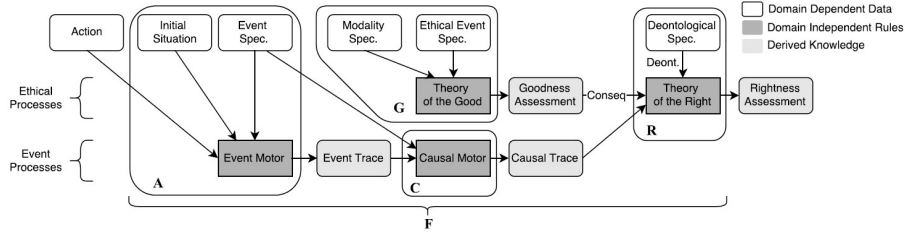


Figure – Une architecture éthique modulaire (Berreby *et al.*, 2017)

Avantages

- Approche générique
- Phase de spécification simplifiée
- Inférence de justification

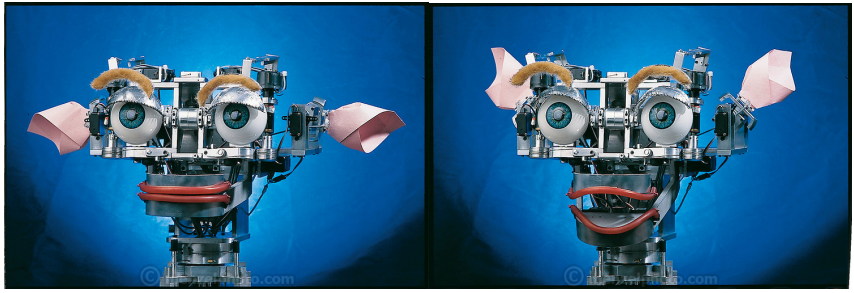
Désavantages

- Complexité du calcul.

Agents émotifs, apparence et confiance humaine

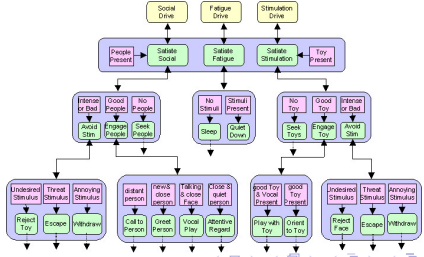
Kismet et les expressions faciales

http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html (photos : copyright Peter Menzel)



Caractéristiques principales

- chaque élément du visage est animé
- détection du ton employé par ses interlocuteurs
- simulation d'une gamme d'émotions



Des robots humanoïdes

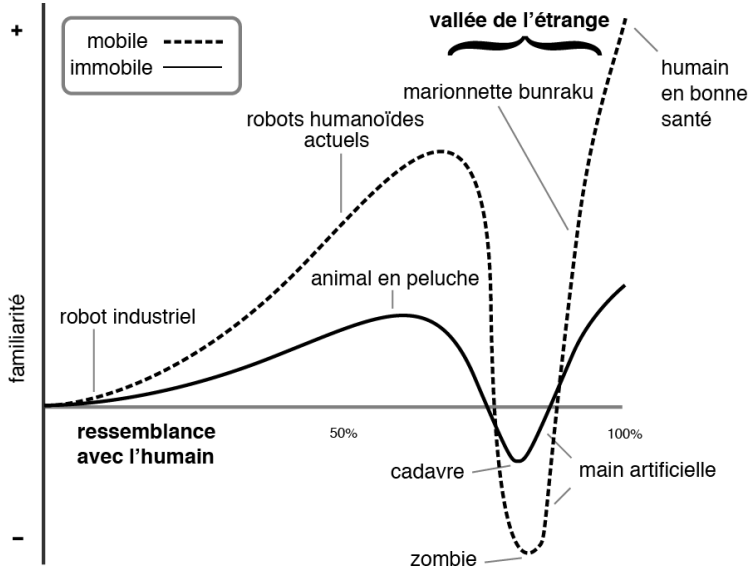
Les Geminoïdes d'Hiroshi Ishiguro (photos : copyrights Hiroshi Ishiguro Laboratory, ATR)

Caractéristiques principales

- machine de taille humaine
- 50 degrés de libertés (tête 13, corps 15, bras et jambes 22)
- système téléopéré
- module conversationel : reconnaissance et synthèse de la parole

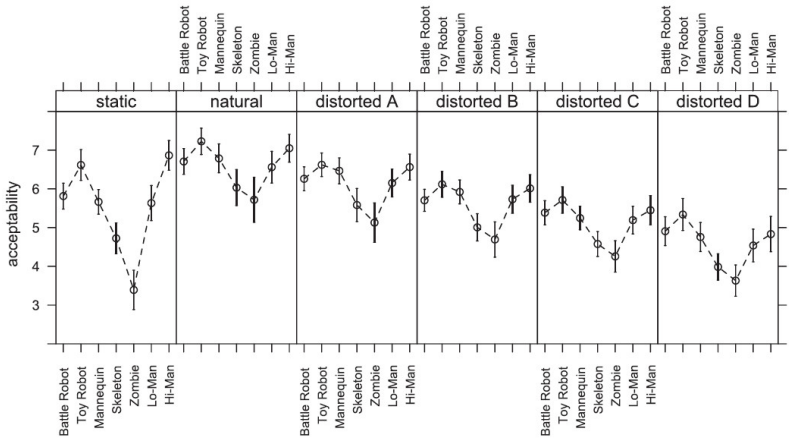
La Vallée de l'étrange (Masahiro Mori)

Robots humanoïdes et sentiment de l'utilisateur



Critiques de la Vallée de l'étrange

La mobilité ne creuse pas la vallée (Piwek *et al.*, 2014)



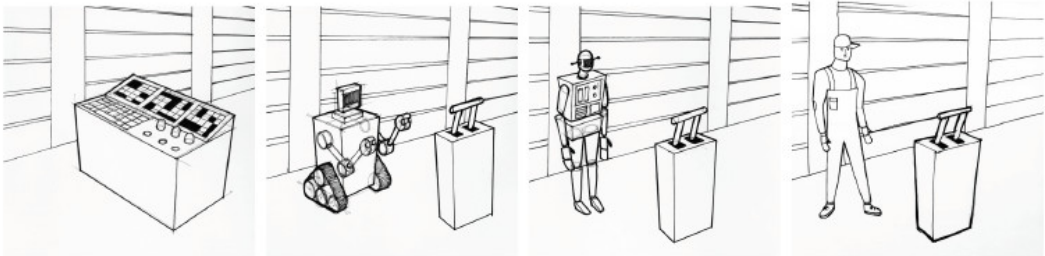
Autres questions

- Quelle persistance temporelle ? La familiarité augmente-t-elle avec des interactions répétées ?
- Quelle valeur locale ? La familiarité exprimée est-elle corrélée à la culture, le genre, l'âge ?

L'apparence des robots change-t-elle nos perceptions de leurs décisions ?

Travaux (2016) de Bertram Malle, professeur de psychologie à la Brown University

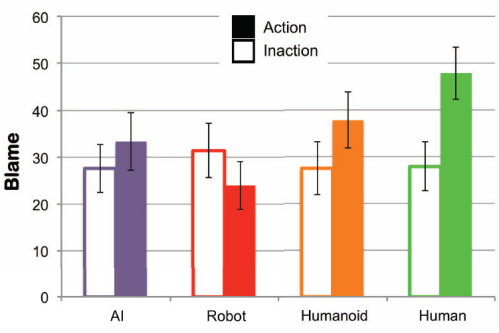
Considérons un dilemme du trolley mettant en scène les acteurs ci-dessous

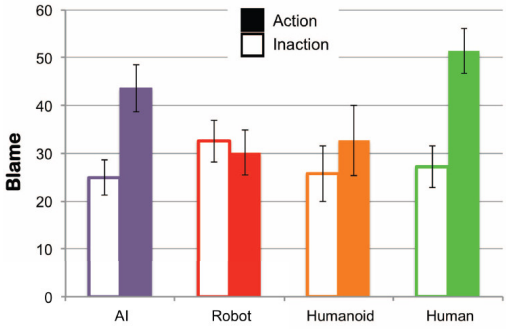


L'apparence des robots change-t-elle nos perceptions de leurs décisions ?

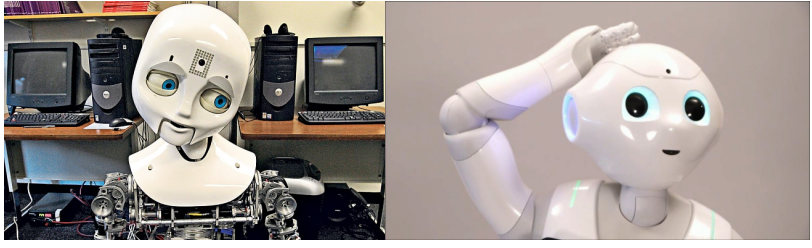
Travaux (2016) de Bertram Malle, professeur de psychologie à la Brown University

633 et 423 participants (quasi-parité homme-femme)





Les humains ont tendance à faire plus confiance aux robots qui expriment des émotions même si leur objectif de vendre des produits ou de gagner à un jeu.

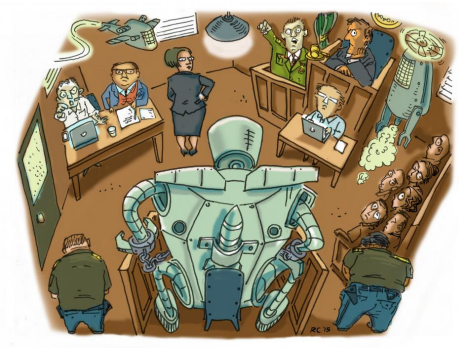


Shim and Arkin (2013), A Taxonomy of Robot Deception and its Benefits in HRI

Droits et robots

Qui est responsable des plans et politiques calculés par un agent autonome ?

(dessin : copyright Christian Russell)



Une question de droit

- Identifier les responsabilités juridiques
- Permettre le dédommagement des victimes
- Conserver une incitation au développement

Que dire des agents artificiels ?

- ce sont des machines et non pas des êtres sensibles
- donc ce sont de outils
- donc ils n'ont pas de droits et de devoirs

Jusqu'à présent, seules les personnes physiques (humains) et les personnes morales (entreprises, associations, états) ont une personnalité judirique

Human Right Watch

- `http://www.stopkillerrobots.org/`
- intégrer un humain à chaque grande étape de décision
- rendre transparent les procéssus et métriques internes des agents artificiels

IEEE Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

- `https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html`
- clarifier les principes légaux et de responsabilité liés aux systèmes intelligents
- intégrer explicitement la notion de valeur dans les processus de décision

Plate-forme Intelligence Artificielle (Caen 2017)

- `https://pfia2017.greyc.fr/`
- 2e Journée Éthique et Intelligence Artificielle

Droits et devoirs des agents autonomes

Personnalité juridique ou morale

Pour $\Longrightarrow$ Charte des droits des robots (Alain Bensoussan)

- Un robot a le droit au respect de sa dignité limitée aux *données à caractère personnel* conservées.
- Conserver une trace des opérations effectuée par un robot pendant une période d'un mois.
- Responsabilité hiérarchique : l'utilisateur, le concepteur, le propriétaire.

Contre $\Longrightarrow$ Règles européennes de droits civils en robotique (Nathalie Nevejans)

- Si un homme reste présent derrière la personne-robot, alors la personnalité robotique est inutile
- Comment une simple machine vide de toute conscience ou volonté pourrait être son propre acteur ?
- Accorde la personnalité à une machine lui donne des droits et des devoirs (notions humaines)

Limites de ce questionnement

- L'humain est le représentant légal de la machine vs. la machine est le mandataire de l'humain
- Il s'agit d'une question de forme qui ne règle pas les détails
- La personnalité morale implique une communauté d'intérêt
- Quel est l'intérêt commun entre les utilisateurs, concepteurs et propriétaires ?

La noxalité dans la Rome antique

- Un *paterfamilias* est responsable des individus et choses placés sous son autorité
- L'abandon noxal, la clé de la noxalité
- La connaissance ou l'ignorance des faits reprochés

Deux fondements : la paternité et la connaissance

La responsabilité des êtres calculants (Alexei Grinbaum)

- responsabilité noxale du programmeur du code-source comme pater de l'individu numérique
- responsabilité matérielle de l'utilisateur-propriétaire du corps physique du robot
- responsabilité noxale de l'individu numérique (et donc du programmeur) pour la collecte de données

Concept intéressant mais inadéquat pour l'époque

- Abandon et système de vengeance
- Responsabilité implicite des choses

Abandon, une option réaliste ?

- pour les concepteurs : un incitatif autrement atteignable, une guerre entre concurrents
- pour les victimes : un calcul compliqué

Idée : paterscientia

- Pater : la responsabilité du fait des animaux
- Scientia : le Code Noir des anciennes colonies françaises

Conclusion

Conclusion

Ce qui faut retenir

Un agent autonome décide de ses actions en fonction des buts qui lui sont donnés

Un agent autonome cherche à réaliser ses buts le plus efficacement possible

Un agent autonome n'a pas de sens commun culturel et empathique

Dans ce cas, être efficace passe parfois par :

- ne pas dire la vérité
- abuser d'une forme de confiance
- ne pas tenir compte de critères éthiques

Sachant que :

- notre jugement change en fonction de l'apparence de l'agent
- l'opacité des décisions peut nuire à une évaluation des responsabilités

Conclusion

Où comment ne pas rester avec un goût amer sur la langue

(Ciorta *et al.*, 2012)

Les utilisateurs ne sont pas toujours conscients de ce qu'ils font et la machine peut leur en faire prendre conscience : un réseau social pro-actif qui prévient les utilisateurs en cas de violation du respect de la vie privée (partage de photos non explicitement autorisé).

(Hamacher *et al.*, 2016)

Les utilisateurs sont à la recherche de transparence : les utilisateurs humains placent plus d'importance à la transparence et au contrôle qu'à l'efficacité des algorithmes.

(Wang *et al.*, 2016)

La collaboration homme-machine est plus efficace : le taux d'erreur dans les diagnostics de cancers est de 3,5% pour les humains et de 7,5% pour les machines, mais que ce taux chute à 0,5% lorsqu'un humain et une machine coopèrent.

Merci de votre attention