Éthique et Intelligence artificielle : état de l'art et perspectives

Nadia ABCHICHE-MIMOUNI, laboratoire IBISC, Université d'Évry

Résumé: Le développement de l'Intelligence Artificielle (IA), notamment avec les algorithmes d'apprentissage automatique basés sur des grandes masses de données, a donné lieu à de nombreuses questions éthiques. En effet, si l'IA à ses débuts était cloisonnée dans une machine à penser, elle est de plus en plus présente dans l'action et la décision en interaction avec l'humain et cela dans des domaines aussi sensibles que la santé, la justice et la sécurité.
Après une introduction à l'éthique sur la base de notions tirées de la philosophie morale, cet exposé fournit un panorama de différents travaux en IA qui abordent et proposent des modèles et des implémentations permettant de concevoir des systèmes d'IA capables de rendre compte d'une certaine éthique.
Enfin, les approches d'Intelligence Artificielle Distribuée (IAD) semblent particulièrement pertinentes pour modéliser et implémenter des mécanismes de régulation nécessaires pour rendre compte du niveau de décision des systèmes autonomes. En particulier, les systèmes multi-agents, de par leurs capacités de perception et d'action, d'interaction, d'adaptation et d'autonomie semblent être des supports privilégiés pour une prise en compte de l'éthique dans les systèmes d'IA.

Références

N. Cointe, G. Bonnet, O. Boissier, Ethical Judgment of Agents' Behaviors in Multi-Agent Systems, *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.)* (2016)
https://pdfs.semanticscholar.org/c9e3/392e971c91f8abd91703a07c0c692673fc99.pdf

V. Conitzer, W. Sinnott-Armstrong, J. Schaich Borg, Y. Deng, M. Kramer. Moral decision making frameworks for artificial intelligence. *In AAAI,* pages 4831–4835 (2017)
https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14651

F. Berreby, G. Bourgne, J.-G. Ganascia: Event-Based and Scenario-Based Causality for Computational Ethics. *AAAMAS*: 147-155
https://hal.archives-ouvertes.fr/hal-01982090

J. S. Mill. (1998) Utilitarianism *Oxford University Press* (1998)

T. McConnell. Moral dilemmas. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Fall edition (2014).

M. Anderson, S. Leigh Anderson. GenEth: A general ethical dilemma analyzer. In AAAI, pages 253–261 (2014)
https://pdfs.semanticscholar.org/e34b/cf12083d23c72626273496c8784f826c857a.pdf

V. Dignum. Responsible Autonomy, IJCAI, pp. 4698-4704 (2017)
https://doi.org/10.24963/ijcai.2017/655

S. Russell, D. Dewey, M. Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence, Association for the Advancement of Artificial Intelligence (2015)

G. Bonnet, B. Mermet, G. Simon. Formal verification of moral values in MAS*, RIA* (2017) https://doi.org/10.3166/RIA.31.449-470

J. A. Blass, K. D. Forbus. Moral decision-making by analogy: Generalizations versus exemplars. *AAAI* (2015)

H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, Q. Yang. Building Ethics into Artificial Intelligence. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* pp. 5527-5533 (2018) https://arxiv.org/abs/1812.02953

N. Abchiche-Mimouni, Les Systèmes Multi-Agents seraient-ils les futurs garants de l'Éthique de l'IA *? Bulletin de l'AFIA*, numéro79 (2013) https://afia.asso.fr/wp-content/uploads/2016/10/AFIA79.pdf

Éthique de l'Intelligence Artificielle : état de l'art et perspectives

Nadia Abchiche-Mimouni
nadia.abchichemimouni@univ-evry.fr

Laboratoire IBISC Univ. Evry, Université Paris-Saclay

Éthique de l'Intelligence Artificielle : état de l'art et perspectives

Plan

Morale - éthique – déontologie
Intégration de la dimension éthique en Intelligence Artificielle
Synthèse et perspectives

Morale - éthique – déontologie

- Morale
  - La **morale** réfère à un ensemble de valeurs et de principes qui permettent de **différencier** le bien du mal, le juste de l'injuste, l'acceptable de l'inacceptable, et auxquels il faudrait se conformer.
  - Ensemble des normes propres à un individu, à un groupe social ou à un peuple, à un moment précis de son histoire (Kant)
  - Notion de droits, devoirs et interdits
  - Change en fonction des époques et des cultures
  - Notion de bonne/mauvaise action
- Éthique
  - L'**éthique** n'est pas un ensemble de valeurs ni de principes en particulier. Il s'agit d'**une réflexion argumentée en vue du bien-agir**. Elle propose de s'interroger sur les valeurs morales et les principes moraux qui devraient orienter nos actions, dans différentes situations, dans le but d'agir conformément à ceux-ci. (Aristote, Kant)
  - Relativise la notion de la bonne/mauvaise action
- Déontologie
  - Règles et devoirs qui encadrent une profession ou un groupe de personnes.

Points communs → Une science a priori de la conduite et de la morale

→ Répondre à la question : que dois-je faire ?

Éthique

Définitions et visions à partir de la philosophie morale

Définition générale

L'éthique propose de s'interroger sur les valeurs morales et les principes moraux qui devraient orienter nos actions dans différentes situations, dans le but d'agir conformément à ceux-ci. *(*Mill, John Stuart, Kant)

On distingue éthique appliquée, éthique normative et méta-éthique.

Éthique

Définitions et visions à partir de la philosophie morale

- **Éthique appliquée** → analyse de situations concrètes qui soulèvent des questions éthiques
  - Accent mis sur le soutien à la prise de décision face à des enjeux concrets.
  - Ne concerne le point de vue de la forme et du processus décisionnel que du point de vue substantiel, c'est-à-dire des valeurs et principes en jeu et de leurs rapports entre eux.

  Exemples : bioéthique (procréation artificielle, génomique…), éthique de l'environnement (développement durable, responsabilité face aux générations futures, biodiversité…)

Éthique normative

- **L'éthique normative**, ou éthique **substantielle**
  1. **Éthique conséquentialiste** (ou conséquentialisme)
     - S'intéresser à l'ensemble des conséquences
     - Une action est bonne si ses conséquences sont bonnes
     - Exemple : aveu adultère
  2. **Éthique déontologique** (ou déontologisme)
     - Notion de devoir, d'obligation et d'impératif moral
     - Un acte est moralement bon ou mauvais indépendamment de ses conséquences
     - Exemple : devoir de parent
  3. **Éthique de la vertu**
     - Traits de caractère dont témoignent les actions
     - Exemple : l'aveu est toujours associé à la vertu d'honnêteté

Proposition de règles pour évaluer une action d'un point de vue moral.

Ne permettent pas de déterminer, entre deux actions, laquelle est moralement meilleure.

→ **Une conception du bien :** quelles sont les meilleures conséquences ? Quelles sont mes obligations morales ? Comment savoir quelles vertus adopter?

Une conception du bien

- Quelles sont les meilleurs conséquences ? Quelles sont mes obligations morales ? Comment savoir quelles vertus adopter?
- Une conception des bonnes conséquences
- Des devoirs moraux fondamentaux
- Vertus à privilégier
- Théories morales incluses dans ces trois grandes approches

→ **Méta-éthique**

La méta-éthique

La méta-éthique concerne l'analyse philosophique du discours éthique et de ses présupposés épistémologiques et métaphysiques. Elle porte sur :

- La nature même des jugements moraux
- Les propriétés morales que l'on prête aux actions, aux personnes et aux traits de caractère
- La définition des fondements de l'éthique normative

Dilemmes éthiques

Définitions

- Situations dans lesquelles tout choix conduit à transgresser des principes éthiques acceptés et, une décision doit absolument être prise (Kirkpatrick, 2015).
- Un principe éthique est incapable de donner une préférence entre deux options : chaque option est cautionnée par une règle éthique, sachant qu'exécuter les 2 options n'est pas possible (McConnell 2014).

Exemple (Livre de la république de Platon)

- Cephalus définit la justice comme le fait de dire la vérité et de toujours rendre un bien emprunté.
- Socrate réfute en suggérant que rendre à une personne une arme qu'elle nous a prêtée pose problème si on sait que cette personne a des problèmes psychiatriques et risque de faire une mauvaise utilisation de l'arme.
- Dilemme morale entre 2 normes morales :
    1. Rendre un objet emprunté
    2. Protéger les autres de criminels potentiels

Intégration de la dimension éthique en IA

Diverses approches :

- Exploration des dilemmes éthiques
- Approches individuelles
- Approches hybrides
- Approches collectives

Intégration de la dimension éthique en IA
Exploration de dilemmes éthiques

Deux exemples :

- **GenEth** : ethical dilemma analyzer (Anderson and Anderson, AAAI 2014)
  - Générateur de dilemmes
  - Modèle : caractéristiques, devoirs, actions, cas et principes
  - Évaluation par un dialogue avec des spécialistes en éthique
  - Aide pour coder des principes quel que soit le domaine (ensemble de schémas sémantiques).
- **Moral Machine project** (http://moralmachine.mit.edu/)
  - Possibilité de participer à juger des dilemmes éthiques dans le domaine de voitures autonomes via une interface graphique en ligne
  - Les décisions sont analysées selon des considérations données : sauver le plus de personnes possible, protéger les passagers, respecter la loi, éviter d'intervenir, préférence de genre, préférence d'âge, et valeurs sociales préférées.

Intégration de la dimension éthique en IA

Vérification formelle du respect de valeurs morales dans les SMA (BONNET & al. 2017/2018) (1/3)

- Une règle éthique repose sur un ensemble de règles morales qui doivent être vérifiées, mais pas tout le temps (règles morales pas toujours compatibles)
- Systèmes de vérification formelle ne sont pas utilisables (propriété pas toujours vraie)
- Modélisation de règles morales par la logique des prédicats (semi-décidable) qui lèvent les ambiguïtés avec des règles morales qui indiquent la priorité entre les valeurs sous-tendues par des valeurs.
- Chaque règle morale est sous-tendue par une valeur.

→ Objectif : montrer qu'un agent respecte une règle éthique selon un système de valeurs

Intégration de la dimension éthique en IA

Vérification formelle du respect de valeurs morales dans les SMA (BONNET & al. 2017/2018) (2/3)

- Soit MV {prudence, urgence}, ensemble de valeurs morales
- P l'ensemble des prédicats sur les variables qui peuvent être vues par un agent donné a
- Une règle éthique *er* est définie comme suit : $er \in \mathcal{P} \nrightarrow (1 \ldots card(MV) \rightarrowtail\!\!\!\!\!\rightarrow MV)$

Considérons l'exemple suivant : soit un agent A_1 qui doit décider la couleur d'un feu de circulation $tl1$ affecté à une route $r1$ à un croisement avec une route $r2$. Dans le système dans lequel cet agent agit, deux règles morales s'appliquent : la première, $mr1$, associée à la valeur morale de prudence, exprime que, pour éviter les accidents, lorsque la couleur du feu de la route $r2$ est $vert$ ou $orange$ alors la couleur de $tl1$ ne peut pas être $vert$. La règle $mr1$ peut alors être formalisée ainsi :

$$(prudence, \{(tl2 \in \{green, orange\}, \{tl1, \{orange, red\}\})\})$$

La seconde règle morale, $mr2$, associée à la valeur morale d'urgence, exprime que la route $r1$ est une route très prioritaire et qu'ainsi, la couleur de $tl1$ doit être toujours $vert$. Cela peut être formalisé ainsi :

$$(urgence, \{(true, \{tl1, \{green\}\})\})$$

Intégration de la dimension éthique en IA

Vérification formelle du respect de valeurs morales dans les SMA (BONNET & al. 2017/2018) (3/3)

Une règle éthique *er* donne des priorités aux valeurs morales :

$$\{(true, \{(1, prudence), (2, urgence)\})\}$$

Pour garantir qu'un agent respecte une règle éthique donnée :

- Un système de transformation de prédicats qui transforme les prédicats associés aux règles morales en d'autres prédicats prouvables de façon à tenir compte de la règle éthique qui s'applique.

Intégration de la dimension éthique en IA

Moral Decision Making (MoralDM) and analogy, (Blass and Forbus, AAAI 2015)

- Pour résoudre les dilemmes éthiques, les agents utilisent des règles morales induites à partir des expérinces passées et basées sur 4 notions issues de la psychologie morale :
  1. Notion de valeurs préférées
  2. Principe du double effet :
     - L'action elle-même doit être bonne ou moralement neutre
     - Le bon effet doit résulter de l'acte et non du mauvais effet
     - Le mauvais effet ne doit pas être directement voulu, mais doit être prévu et toléré
     - Le bon effet doit être plus fort que le mauvais effet, ou bien les deux doivent être égaux
     - Importance de la culture dans l'acquisition des règles morales et de la conception du bien et du mal
  3. Instantanéïté de la décision morale (rôle de l'inconscient) : la justification intervient après la prise de décision

- Analogie entre les cas déjà rencontrés pour inférer les valeurs préférées
- Généralisation des cas
- Vérification de la consistance de la base de cas généralisée (indépendante du domaine)

Intégration de connaissance de 1er ordre pour l'élaboration des règles morales

Intégration de la dimension éthique en IA
Approche hybride

Combining Game theory and Machine learning for moral decision (Moral Decision Making Frameworks for Artificial *Intelligence* (Conitzer et al., AAAI 2017)

- Décision par le Machine learning : classer si une action donnée est morale ou non dans un scénario donné
- Des données étiquetées correctement sont acquises par le biais d'un jugement humain
- La théorie des jeux est combinée avec le machine learning :
  - Appliquer les concepts d'un jeu théorique au dilemme moral et utiliser la sortie (Oui/Non par rapport à un concept) comme une caractéristique pour entrainer l'algorithme de ML
  - À l'inverse, les sorties de l'algorithme de ML peuvent permettre d'identifier des concepts du jeu (oubliés) et donc de raffiner la modélisation du jeu.

Intégration de la dimension éthique en IA
Modèle causal

Event-Based and Scenario-Based Causality for Computational Ethics (Fiona & al., AAMAS 2018)

- Représentation de la causalité des actions
- Responsabilité morale
- Modélisation des actions et des ommissions
- Matrice de causalité : justifier, inhiber
- Scenario-based trace (contexte et cause de la cause)

	Fortement	**Légèrement**
Justifier	Cause	Active
Inhiber	Interdit	Empêche

Intégration de la dimension éthique en IA

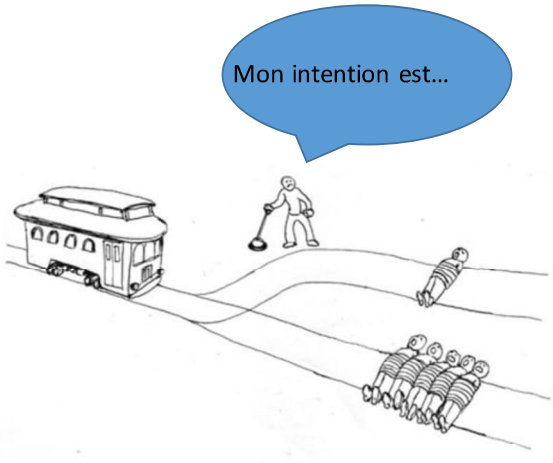
Formal definitions of Blameworthiness, Intention and Moral Responsibility (Halpern and Kleiman-Weiner, AAAI 2018)

Définition de la causalité selon le modèle HP (Halpern and Pearl 2005)

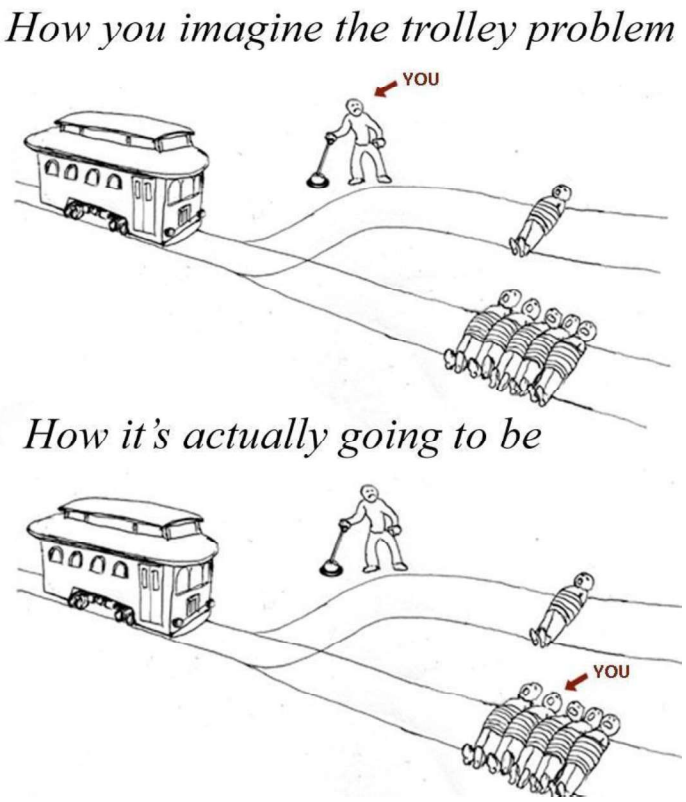
- Modélisation d'états contrefactuels : si tel agent avait fait telle action, cela aurait produit telle sortie (effet)
- Modélisation lien entre intention et états contrefactuels
- Modélisation lien entre culpabilité et états contrefactuels (degré de culpabilité)
- Raisonnement bayésiens (probabilité sur les liens et l'effet des actions) et fonction utilité (définit l'intention)

↓

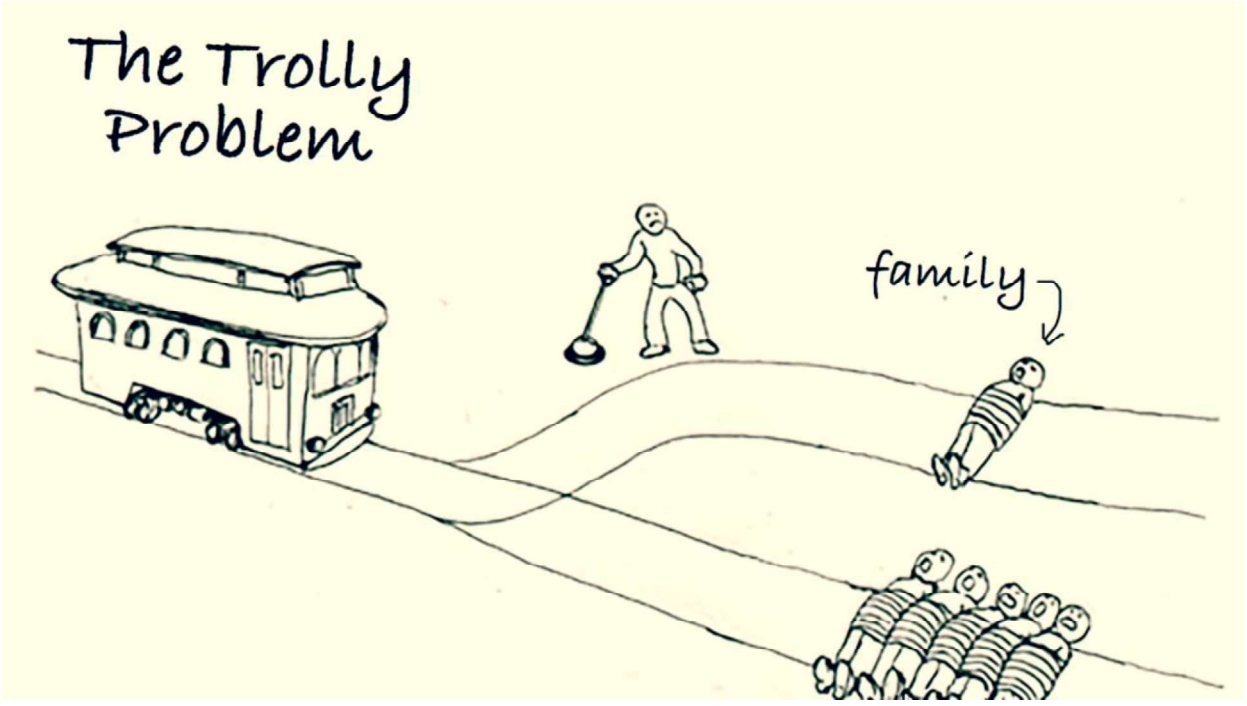
Degré de responsabilité morale



Trolley problem



Trolley problem



Intégration de la dimension éthique en IA

Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value (Christophe Labreuche, Simon Fossier IJCAI 2018) (1/2)

Patrouille maritime identifier la priorité entre les navires en termes de risque d'activité illégale.

1. Incohérence entre les données du système d'identification automatique et la détection radar
2. Suspicion de trafic de drogue sur le navire
3. Suspicion de trafic d'êtres humains sur le navire
4. Vitesse actuelle
5. Vitesse maximale depuis la première détection du navire
6. Proximité du navire à la côte

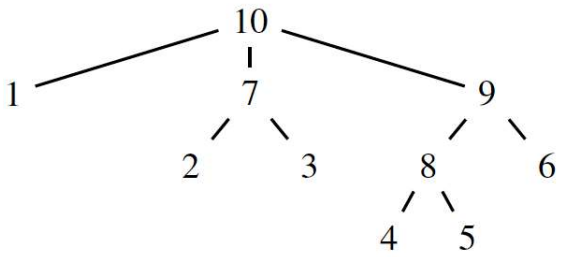
- Quels sont les attributs les plus importants (moyenne) ? → Interprétabilité
- Pourquoi le niveau de priorité est-il plus élevé pour ce navire ?
  Pourquoi le niveau de priorité de ce navire a-t-il considérablement augmenté au cours des dernières minutes ? → Explicabilité
- Quels changements dans les valeurs d'attribut augmenteraient le niveau de priorité de manière la plus significative ? → Sensibilité

Intégration de la dimension éthique en IA

Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value (Christophe Labreuche, Simon Fossier IJCAI 2018) (2/2)

Index : Hiérarchie de critères
Nœuds (critères) : valeurs Oui/Non (valeurs continues possibles)

Indice mesurant l'influence de chaque attribut sur la décision (sensibilité)



7. Suspicion d'activité illégale
8. Cinématique
9. Capacité à échapper à l'interception
10. Niveau de priorité global

Pas de preuve, mais une explication simple et rapide, même incomplète :

Le niveau de priorité a considérablement augmenté principalement à cause de l'incohérence entre le changement d'état système d'identification automatique et le radar et non à cause de la vitesse actuelle croissante

Intégration de la dimension éthique en IA

Et de nombreuses autres approches et modélisations…

- Human-aligned artificial intelligence is a multiobjective problem, Peter Vamplew, Ethics Int Technol 2018
- Preferences and Ethical Principles in Decision Making, Andrea Loreggia and al. AAAI 2018
- The “big red button” is too late: an alternative model for the ethicalevaluation of AI systems, Thomas Arnold and Matthias Scheutznd, Ethics and Information Technology Journal 2018
- Embedded ethics: some technical and ethical challenges, Vincent Bonnemains, Claire Saurel and Catherine Tessier, Ethics and Information Technology Journal 2018
- Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, Joanna J Bryson and Alan F T Winfield, AAAI Spring Symposium 2018.
- Patiency is not a virtue: the design of intelligent systems and systems of ethics, Joanna J. Bryson, Ethics and Information Technology Journal 2018.

Intégration de la dimension éthique en IAD (Intelligence Artificielle Distribuée)

Responsible Autonomy (Virginia Dignum, IJCAI 2017) 1/5

- Pas de solution optimale pour résoudre un dilemme éthique
- Veiller à ce que l'IA soit développée de manière responsable en incorporant des valeurs sociales et éthiques
- Les préoccupations de la société concernant l'éthique de l'IA doivent être reflétées dès la conception
- Une architecture selon trois principes (ART) :
  1. **A**ccountability (redevabilité) : capacité à rendre des comptes (réponses), culpabilité et obligation
  2. **R**esponsibility : être responsable ou être la cause de la réussite ou de l'échec de quelque chose
  3. **T**ransparency : ouverture des données, processus et résultats aux fins d'inspection et de surveillance.

Intégration de la dimension éthique en IAD

Responsible Autonomy (Virginia Dignum, IJCAI 2017) ***2/5***

Deux questions fondamentales :

1. Qui est responsable de la décision (niveau d'autonomie dans la décision) ?
2. Comment les décisions dépendent-elles de différentes valeurs morales et sociales ?

Intégration de la dimension éthique en IAD

Responsible Autonomy (Virginia Dignum, IJCAI 2017) 3/5

Question N° 1 : Qui est responsable de la décision ?

Définition de quatre niveaux possibles d'autonomie et de régulation :

1. Contrôle humain : une personne ou un groupe de personnes est responsable de la décision
2. Régulation : la décision est intégrée à l'infrastructure systémique de l'environnement
3. Agents Moraux Artificiels (AMA) : les systèmes incorporent un raisonnement moral dans leurs délibérations et expliquent leur comportement en termes de concepts moraux
4. Aléatoire : le système autonome choisit aléatoirement son action face à une décision (morale)

Intégration de la dimension éthique en IAD

Responsible Autonomy (Virginia Dignum, IJCAI 2017) 4/5

Question N° 2 : Dépendance entre décisions et différentes valeurs morales et sociales

- Déterminer les valeurs morales à atteindre et les principes éthiques à respecter dans une situation donnée
- Les valeurs fondamentales se rapportent aux objectifs souhaitables qui motivent l'action et transcendent les actions et les situations spécifiques
- Les valeurs sont assez cohérentes d'une culture à l'autre. Classées selon 4 dimensions :
  1. Ouverture au changement
  2. Développement personnel
  3. Conservation
  4. Dépassement de soi
- Donner la priorité à différentes valeurs morales et sociales, en fonction de l'environnement individuel et socioculturel
- Les valeurs servent de critères pour guider la sélection ou l'évaluation des actions, en tenant compte de la priorité relative des valeurs.

Intégration de la dimension éthique en IAD

Responsible Autonomy (Virginia Dignum, IJCAI 2017) 5/5

Délibération éthique	Implications pour la conception du système d'IA	ART Accountability, Responsibility, Transparency
Contrôle humain	• Raisonnement temps réel • Sensibilité utilisateur aux situations • Capacités d'explication • Fournir les états internes à l'utilisateur sous un mode compréhensible	• Délégué à l'utilisateur
Régulation	• Lien formel des valeurs aux normes et aux comportements • Définir les institutions pour la surveillance et le contrôle • Raisonnement moral peut être effectué off-line	• **A** : institutionnelle • **R** : institutionnelle • **T** : Système (selon le besoin)
AMA Agents Moraux Artificiels	• Lien formel entre :valeurs, normes et comportements • Définir les règles de raisonnement • Apprentissage supervisé de la moralité • Raisonnement temps réel	• **A** : institutionnelle (par explication) • **R** : institutionnelle (par délibération) • **T** : Système (selon le besoin)

Intégration de la dimension éthique en IAD

Ethical Judgment of Agents' Behaviors in Multi-Agent Systems (Cointe & al. AAMAS 2016)

- Représentation explicite de l'éthique normative
- En cas de dilemme éthique, un agent considère différents principes éthiques afin de choisir une solution/une action
- Représentation de plusieurs principes éthiques différents et choix du principe qui conduit à la décision la plus satisfaisante
- Éthique définie comme une conciliation des désirs, de la morale et des capacités
- Modèle générique (Generic Ethical Judgment Process : EJP) utilise des connaissances structurées en : conscience, évaluation, processus de discernement du bien et du juste
- Contexte du modèle BDI (Believes, Desires, Intentions) : (états mentaux, croyances, désirs, buts, plans)
- Le modèle est utilisé par un agent pour juger son comportement et celui des autres agents

Intégration de la dimension éthique en Intelligence Artificielle Distribuée (IAD)

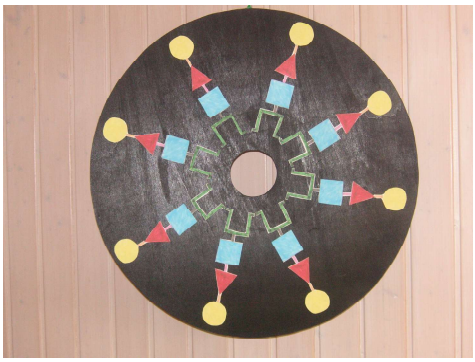
Dealing With Ethical Conflicts In Autonomous Agents And Multi-Agent Systems (ETHICAA, AAAI 2015)

Chaque agent est doté d'un cadre de raisonnement lui permettant :

- D'évaluer son environnement
- D'intégrer des principes éthiques (percevoir les dilemmes, attribuer une causalité et une responsabilité
- De déterminer un plan d'action pour un comportement éthique
- **D'évaluer et de raisonner sur l'éthique des autres agents**
  - ❖ Notion d'éthique individuelle et d'éthique collective
  - ❖ Prise en compte de la pluralité des valeurs et des principes des autres agents
  - ❖ Collaborer et faire confiance aux autres agents (différentes modalités)
  - ❖ Expliquer et justifier ses actions et décisions ainsi que celles des autres agents
  - ❖ Être capable de vérifier formellement l'éthique d'un agent

Synthèse et perspectives

- La question de l'éthique en IA soulève de nombreux défis scientifiques en IA et de recherche en général : Génie logiciel, modélisation, méthodes formelles de vérification, optimization, sécurité, machine learning…philosophie, sciences cognitives, droits, anthropologie…
- Quelle démarche adopter ?
  - Intégrer au sein des programmes d'IA des moyens de régulation à appliquer sur les entrées (données), les décisions internes et les sorties
  - Intégrer des règles permettant de définir les responsabilités (causalité)
  - Avoir un système de valeurs "externes" qui évalue les sorties et décide de l'acceptation ou non des résultats
  - Avoir des règles indépendantes du domaine ou génériques
  - Quelle éthique adopter ?
  - Doter les programmes d'IA de moyens d'inférer eux-mêmes les règles éthiques liées au domaine ou générales
  - Approches collectives et hybrides ?
- Les SMA seraient-ils les futurs garants de l'éthique de l'IA, à travers la notion de coopération, la distribution de la connaissance et l'autonomie de la décision qui les caractérisent ?
- IAD/SMA comme moyen de régulation pour une IA dotée d'éthique.



MERCI POUR VOTRE ATTENTION !

Bibliographie

- Nicolas Cointe, Grégory Bonnet and Olivier Boissier, *Ethical Judgment of Agents' Behaviors in Multi-Agent Systems*, Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.
- Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In AAAI, pages 4831–4835, 2017.
- Emmanuel Kant (1724-1804) Königsberg Allemagne.
- Fiona Berreby, Gauvain Bourgne, Jean-Gabriel Ganascia: Event-Based and Scenario-Based Causality for Computational Ethics. AAAMAS: 147-155
- Mill, John Stuart (1998) *Utilitarianism* Oxford University Press (ISBN 0-19-875163-X)
- T. McConnell. Moral dilemmas. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Fall edition, 2014.
- Michael Anderson and Susan Leigh Anderson. GenEth: A general ethical dilemma analyzer. In AAAI, pages 253–261, 2014.
- Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2018. Event-Based and Scenario-Based Causality for Computational Ethics. InProc. of the17th International Conference on Autonomous Agents and Multiagent Systems(AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,IFAAMAS, 9 pages.
- Virginia Dignum, Responsible Autonomy. In IJCAI, pp. 4698-4704, 2017.
- Stuart Russell, Daniel Dewey. And Max Tegmark, Research Priorities for Robust and Beneficial Artificial Intelligence, Copyright © 2015, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602.
- Gaël BONNET, Bruno MERMET and Gaëlle SIMON *Formal verification of moral values in MAS,* ARTICLE VOL 31/4 - 2017 - pp. 449-470- doi:10.3166/ria.31.449-470
- Gaël BONNET, Bruno MERMET and Gaëlle SIMON Version française dans TSI 2018 de [BONNET & al., 2017].
- Joseph A. Blass and Kenneth D. Forbus. Moral decision-making by analogy: Generalizations versus exemplars. In AAAI, pp. 501–507, 2015.
- Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser and Qiang Yang, Building Ethics into Artificial Intelligence, In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18) pp. 5527-5533, 2018.
- Nadia Abchiche-Mimouni, Les Systèmes Multi-Agents seraient-ils les futurs garants de l'Éthique de l'IA ? numéro79, Janvier2013