Journée Société et IA

Frédéric Alexandre, Grégory Bonnet, Ikram Chraibi Kaadoud, Jean-Gabriel Ganascia

AfIA

Association française
pour l'Intelligence Artificielle

SIA

Journée Société & IA

PFIA 2024

Table des matières

Éditorial

Journée Société & IA

Le groupe de travail ACE (Aspects Computationnels de l'Éthique) du GDR RADIA et Inria Bordeaux ont organisé la journée « Société et IA » les 1er et 2 juillet 2024 dans le cadre de la Plate-Forme Intelligence Artificielle de l'AFIA (PFIA 2024).

Depuis plusieurs années, des comités réunis à l'initiative d'université, d'États, de puissances supra-étatiques comme la Commission Européenne, de sociétés savantes ou d'organisation non gouvernementales réfléchissent aux questions d'éthique de l'Intelligence Artificielle et à sa régulation. Ces réflexions ont abouti entre autres sur la notion de systèmes informatiques dignes de confiance qui sont à mettre en perspective avec les problématiques en éthique artificielle.

Cette journée avait pour objectif de réunir les communautés travaillant sur l'Intelligence Artificielle de confiance, l'éthique artificielle et plus généralement sur tout ce qui est en lien avec l'impact social de l'Intelligence Artificielle. Dans une volonté d'ouverture tant aux communautés de recherche travaillant déjà sur ces problématiques, qu'aux non-spécialistes intéressés, nous avons encouragé toutes les contributions relatives à ces sujets, qu'elles portent sur les aspects techniques, juridiques, philosophiques ou sociologiques de l'Intelligence Artificielle ou sur les impacts industriels de son déploiement.

Sans être exhaustif, les thématiques de cette journée étaient :
— les systèmes informatiques dignes de confiance
— l'interaction humain–machine de confiance
— la réglementation des systèmes d'intelligence artificielle
— l'intelligence artificielle frugale
— la modélisation du raisonnement éthique
— la prise de décision juste et équitable
— le respect de la confidentialité
— la transparence et la lisibilité des décisions automatisées
— l'explicabilité des décisions automatisées
— la prévention des biais et de la discrimination

Dans la littérature, une intelligence artificielle de confiance, ou dite digne de confiance, doit satisfaire au moins trois propriétés : elle doit être robuste, licite et éthique. C'est pour cela que les journées étaient structurées en trois sessions principales, portant respectivement sur les questions d'explicabilité et de robustesse aux biais, sur les questions de réglementation et d'usage, et enfin sur les questions d'éthique computationnelle. Chaque session débutait par une conférence invitée sur le sujet (respectivement données par Laurent Simon, Nathalie Nevejans et Cédric Brun, puis était suivie de présentations des contributions qui avaient été soumises. Tout au long des deux jours, chaque session a vu un peu plus d'une quarantaine de participants.

Frédéric Alexandre, Grégory Bonnet, Ikram Chraibi Kaadoud, Jean-Gabriel Ganascia

Contributeurs

Comité d'organisation

— Frédéric Alexandre (Inria Bordeaux, Bordeaux INP)
— Grégory Bonnet (GREYC, Université de Caen)
— Ikram Chraibi Kaadoud (Inria Bordeaux, Université de Bordeaux)
— Jean-Gabriel Ganascia (LIP6, Sorbonne Université)

Contributeurs

— Frédéric Alexandre (Inria Bordeaux, Bordeaux INP)
— Grégory Bonnet (GREYC, Université de Caen)
— Nadjet Bourdache (GREYC, Université de Caen)
— Gauvain Bourgne (LIP6, Sorbonne Université)
— Cédric Brun (SPH, Université Bordeaux Montaigne)
— Florent Castagnino (LEMNA, IMT Atlantique)
— Ikram Chraibi Kaadoud (Inria Bordeaux, Université de Bordeaux)
— Matthieu Delahaye (Lab-STICC, IMT Atlantique)
— Lina Fahed (Lab-STICC, IMT Atlantique)
— Guillaume Gervois (LIP6, Sorbonne Université)
— Mélanie Gornet (Institut Polytechnique de Paris)
— Clara Fontaine-Say (Chercheuse indépendante)
— Jeanine Harb (Beink Dream)
— Hélène Herman (Centre d'Etude des Mouvements Sociaux)
— Philippe Lenca (Lab-STICC, IMT Atlantique)
— Jeanne Le Peillet (Beink Dream)
— Marie-Jean Lesot (LIP6, Sorbonne Université)
— Alice Maranne (ENSEIRB-MATMECA, Talence)
— Fabrice Muhlenbach (Laboratoire Hubert Curien, Université Jean Monnet)
— Abdel-Illah Mouaddib (GREYC, Université de Caen)
— Nathalie Nevejans (DEP, Université d'Artois)
— Baptiste Pesquet (Inria Bordeaux, Bordeaux INP)
— Laurent Simon (LaBRI, Université de Bordeaux, Bordeaux INP)
— Mihail Stojanovski (GREYC, Université de Caen)
— Sarra Tajouri (LAMASADE, Université Paris-Dauphine PSL)
— Raphael Teitgen (Beink Dream)
— Thierno Tounkara (LITEM, Université d'Evry / IMT-BS, Université Paris-Saclay)
— Alexis Tsoukiàs (LAMASADE, Université Paris-Dauphine PSL)
— Robert Voyer (LITEM, Univ. d'Evry / IMT-BS, Univ. Paris-Saclay / LASCO, Univ. Paris Descartes)

Conférences invitées

Quel sens donner à l'IA de confiance ?

Cédric Brun

Université Bordeaux Montaigne, Département de philosophie
Institut des Maladies Neurodégénératives, UMR 5392, CNRS-Université de Bordeaux
Équipe Neurosciences, Humanités et Société

Cedric.Brun@u-bordeaux-montaigne.fr

Résumé

La notion d'IA de confiance est critiquée comme un cas d'erreur de catégorie (la confiance ne pourrait concerner que les relations entre sujets humains) qui ne peut être évitée que si l'on emploie le terme confiance comme une métaphore ou si l'on réduit la confiance à la fiabilité. Dans cette intervention, nous proposons de retracer les grandes lignes des arguments philosophiques qui ont été formulés contre et pour la notion d'IA de confiance avant de proposer un cadre permettant d'espérer parvenir à aligner les systèmes d'IA avec les valeurs humaines.

Biographie

Maître de conférences au département de philosophie de l'Université Bordeaux Montaigne, dont il a été le directeur entre 2015 et 2018, il enseigne la philosophie générale des sciences, la philosophie de l'esprit, la philosophie des sciences cognitives et des neurosciences et l'histoire de la philosophie de langue anglaise en master ainsi que la philosophie des sciences, la logique, la philosophie de l'esprit et l'histoire de la philosophie de langue anglaise en licence. Jusqu'en 2011, ses recherches se sont concentrées sur l'histoire de la philosophie classique de langue anglaise et sur la philosophie de l'esprit. Depuis 2011, il a orienté son activité de recherche vers la philosophie des neurosciences. Ses recherches actuelles portent sur 1) la question du réductionnisme et de l'antiréductionnisme en philosophie générale des sciences et en philosophie des neurosciences, 2) la question de la nature de l'explication en neurosciences, et 3) la question des conditions épistémologiques et métaphysiques pour l'intégration des différents niveaux d'exploration du système nerveux central dans les neurosciences cognitives et comportementales, à travers la définition de normes de contraintes sur les modèles (computationnels et expérimentaux).

La réglementation de l'intelligence artificielle dans l'Union européenne

Nathalie Nevejans

Université d'Artois
Centre Droit Éthique et Procédures (UR 2471)
Chaire IA Responsable (ANR-19-CHIA-0008)

nathalie.nevejans@univ-artois.fr

Résumé

Face aux risques potentiels que pourraient causer certains usages de l'intelligence artificielle (IA), l'Union européenne a choisi la voie de la réglementation. Le Règlement sur l'IA, appelé *AI Act*, devrait être définitivement adopté pour l'été 2024. Ce sera la première législation au monde destinée à encadrer largement la mise sur le marché ou en service, ou encore l'utilisation des systèmes d'IA, y compris les modèles et systèmes GPAI, par les fournisseurs et les déployeurs. L'*AI Act* introduit une classification des systèmes d'IA selon les risques qu'ils peuvent engendrer et instaure une gradation des contraintes légales pour leur mise en conformité en fonction de leur niveau de risque, depuis le risque minimal jusqu'aux usages totalement interdits. La violation de ces règles sera sanctionnée par des amendes administratives importantes. Bien que l'*AI Act* constitue une avancée importante pour la sécurité juridique des entreprises et la protection des personnes, susceptible d'inspirer le reste du monde, ce texte n'est cependant pas dénué de défauts.

Biographie

Nathalie Nevejans est professeure de droit à l'Université d'Artois (France). Spécialisée en droit et éthique de la robotique et de l'IA, elle est titulaire de la Chaire IA Responsable (ANR-19-CHIA-0008). Autrice de nombreux articles, elle participe à des événements nationaux et internationaux provenant tant du monde académique que de divers secteurs professionnels (industrie, santé, assurance, etc.) Elle est aussi régulièrement auditionnée en tant qu'experte auprès de diverses instances européennes sur le droit et d'éthique de l'IA. Elle est membre du Centre Droit, Éthique et Procédures (UR 2471). Son livre « Traité de droit et d'éthique de la robotique civile », LEH éditions (1232 pages), paru en 2017, a été récompensé par le prix Francis Durieux 2019 de l'Académie des sciences morales et politiques.

Un besoin de confiance artificielle pour l'intelligence artificielle

Laurent Simon

LaBRI
ENSEIRB-MATMECA
Bordeaux INP

lsimon@labri.fr

Résumé

L'Intelligence Artificielle (IA) est au contact d'un nombre croissant de nos activités, suivant de peu la numérisation de tous les pans de nos vies, et ne se limitant plus au travail. Ses applications se réinventent et ouvrent sans cesser de nouveaux horizons dans une course aux performances. Cependant, si les progrès sont tout à fait impressionnants, des questions se posent sur leurs limites et les garanties que l'on pourrait en attendre. La course à la performance s'est affranchie de fondations éthiques et de confiance, qui sont devenues des problèmes annexes, à traiter dans un second temps. Pourtant, rien n'indique que cela soit simplement possible. Dans cet exposé, nous présenterons les raisons pour lesquelles nous devons être réalistes et exigeants sur la nécessaire confiance que nous devons construire autour de ces systèmes, ainsi que montreront en quoi celle-ci est malheureusement orthogonale aux progrès réalisés.

Biographie

Enseignant-chercheur en Intelligence Artificielle à Bordeaux INP et à l'Université de Bordeaux, Laurent Simon est Professeur des Universités à l'ENSEIRB-MATMECA (Bordeaux INP). Il effectue sa recherche au LaBRI (Laboratoire Bordelais de Recherche en Informatique) autour de la conception d'outils de résolution de problèmes logiques fortement combinatoires, typiquement utilisés dans des problématiques de preuves ou de certifications, qu'il applique aux problématiques d'IA hybrides. Il a été porteur de deux projets ANR sur ce sujet et porte depuis 2022 une chaire Industrielle autour de l'IA digne de Confiance. Il est le directeur du département informatique de l'ENSEIRB-MATMECA et président de l'association française de programmation par contraintes, éditeur associé de JAIR, l'un des principaux journaux internationaux en IA. Il enseigne l'IA depuis plus de 20 ans.

Session 1 : confiance & explicabilité

Détection de biais et intégration de connaissances expertes pour l'explicabilité en IA

Matthieu Delahaye[1], Lina Fahed[1], Florent Castagnino[2], Philippe Lenca[1]

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
[2] IMT Atlantique, LEMNA, F-44307 Nantes, France

prénom.nom@imt-atlantique.fr

Le déploiement de modèles d'apprentissage automatique « boîtes noires », dans des secteurs sensibles notamment, a entraîné un fort besoin d'explicabilité adapté au niveau de compréhension des acteurs décideurs. Dans le secteur sensible de la sécurité urbaine, de nombreux modèles de prédiction d'infractions (contraventions, délits, crimes) dans le temps et l'espace ont été proposés. En France, la plupart de ces modèles ne sont plus utilisés à cause, principalement, de l'absence d'informations complémentaires à l'intuition des policiers [2]. Dans l'objectif de comprendre les phénomènes liés à la sécurité urbaine et pour apporter des informations pertinentes aux policiers, nous souhaitons proposer un nouveau modèle d'Intelligence Artificielle eXplicable (XAI) [3] qui relève les deux défis suivants :

Défi 1 : Détection de biais par l'explication. Le biais cognitif présent dans le jugement et la décision humaine est un phénomène naturel [1]. Sachant qu'un modèle d'apprentissage automatique, par définition, extrait des connaissances à partir de données qui sont, elles, souvent générées et collectées par l'humain, cela engendre la présence de biais dans le modèle. La littérature s'est principalement concentrée sur l'atténuation des biais algorithmiques, ce qui peut entraîner une incapacité de justification du résultat du modèle [1]. Or, dans un contexte d'aide à la décision à fort impact social, il est nécessaire de considérer les acteurs impliqués et leurs contextes sociaux tout au long de la modélisation afin de conserver un modèle fiable et non biaisé. Notre intérêt se porte plutôt sur les étapes préliminaires au traitement de biais : la détection et l'évaluation des biais. Pour détecter exhaustivement les biais du modèle, nous cherchons à expliquer son comportement. L'explication des modèles peut se faire aussi bien à l'échelle locale que globale [3]. Localement, générer des contre-factuels, par exemple, permet de modifier légèrement les instances de façon à changer la prédiction. Dans ce cas, la détection de biais peut se faire en analysant des cas précis. À l'échelle globale, l'explication est donnée à propos du comportement entier du modèle ce qui permet de détecter les biais de façon plus complète. Dans notre contexte, les techniques d'explication liées à ces deux échelles sont à explorer.

Défi 2 : Généralisation par intégration de connaissances. Les modèles d'apprentissage automatique fondés exclusivement sur des données, parfois imparfaites, sont limités en terme de généralisation [5]. Dans le but de préciser leur raisonnement, nous proposons d'intégrer des connaissances expertes durant l'apprentissage du modèle afin qu'il assimile le contexte et les contraintes extérieures n'étant pas explicitement dans les données [5]. Dans notre projet, les connaissances expertes sont le résultat d'entretiens réalisés auprès des professionnels de la sécurité, des criminologues et sociologues de la police. La pluralité d'acteurs limite les biais induits par la subjectivité de chaque expert. La piste envisagée est d'incorporer les connaissances dans la fonction de coût du modèle [5] qui se verra ajouter un nouveau terme pouvant être vu comme un terme de régularisation permettant, par exemple, de tempérer la stigmatisation. Afin de vérifier la fiabilité des explications, l'approche d'eXplonatory Interactive Learning (XIL) [4], qui fait intervenir un expert donnant un retour sur la conformité des explications fournies, sera considérée. L'intervention experte contribue à rendre le modèle plus réaliste et renforce la relation de confiance de la police envers l'outil.

Références

[1] Salem Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 11 :6271, 07 2021.

[2] La Quadrature du Net. La police prédictive en France : contre l'opacité et les discriminations, la nécessité d'une interdiction. `https://www.laquadrature.net/`, Janvier 2024.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5) :1–42, 2018.

[4] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat Mach Intell*, 2(8) :476–486, 2020.

[5] L. von Rüden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 35(1) :614–633, 2023.

Modéliser la confiance d'un agent décisionnel

Baptiste Pesquet, Frédéric Alexandre

Centre INRIA de l'Université de Bordeaux, CNRS, Bordeaux INP

baptiste.pesquet@inria.fr

Contexte et objectif

La prise de décision est un phénomène cognitif bien étudié et différents cadres de modélisation permettent de créer des agents décisionnels artificiels reproduisant fidélement certaines caractéristiques de la décision humaine. Estimer la confiance qu'un agent a dans sa décision est une faculté métacognitive et, à ce titre, elle peut le conduire à modifier son comportement. La confiance est fréquemment évoquée dans le développement de l'IA moderne mais ses caractéristiques sont cependant beaucoup moins bien connues. En nous reposant sur différents domaines d'étude, nous cherchons à proposer un cadre de modélisation pertinent de la confiance ainsi que des agents artificiels dotés de cette capacité métacognitive. Nous visons ainsi l'augmentation de leurs performances, mais également de leur explicabilité et de leur acceptabilité.

Modélisation cognitive

Prise de décision

Une décision peut être décrite comme un processus d'accumulation d'indices (*evidence*) issus de notre perception ou de valeurs apprises. Un autre élément-clé est le temps de réaction associé, décrit par un seuil que doit atteindre cette accumulation. Un ensemble de modèles repose sur ce principe [1]. Le plus connu d'entre eux est le *Diffusion Decision Model* (DDM). Conçu pour les choix binaires, ce modèle exploite un seul accumulateur pour représenter la dynamique d'intégration des indices vers l'un des deux seuils de décision. D'autres modèles de cette famille utilisent plusieurs accumulateurs, soit indépendants et associés chacun à un choix, soit en compétition selon différents mécanismes (*best-vs-next*, pondération, inhibition mutuelle, etc).
Par ailleurs, des modèles plus récents étudient des situations dans lesquelles les décisions sont suivies de récompenses qui peuvent, via apprentissage par renforcement, modifier les décisions à venir [2].

Confiance

Comme processus métacognitif, la confiance a deux volets : (1) évaluer la qualité de sa décision permet d'estimer son niveau de confiance pour ensuite (2) adapter éventuellement son comportement, selon ce niveau de confiance. Certaines approches envisagent l'estimation de la confiance comme un processus post-décisionnel basé sur le même principe d'accumulation que le DDM [3]. Suite à une prise de décision permise par une accumulation d'indices, on poursuit cette accumulation pour voir si cela infirme ou confirme cette décision et, selon cette dérive, cela permet d'estimer le niveau de confiance accordé à cette décision.

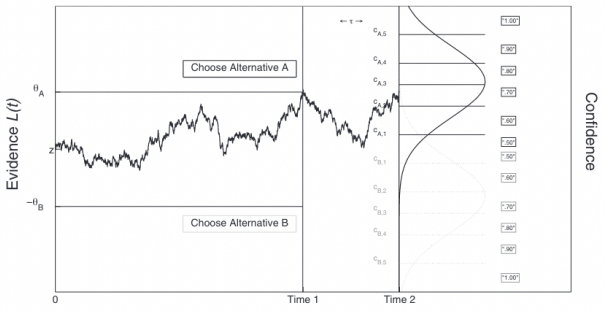

FIGURE 1 – Illustration de l'accumulation d'indices pour la décision et la confiance, extrait de [3]

Approche développée

Sur la base de ces études, nous développons un agent artificiel associant apprentissage et prise de décision, capable de choix non binaires et dont le niveau de confiance dans ses décisions est estimé à l'aide d'un modèle à accumulateurs en compétition. Cet agent, conçu pour agir dans son environnement, pourra modifier son comportement sur la base de cette estimation. Nous discuterons également la possibilité d'utiliser ces évaluations pour offrir des garanties permettant de proposer un modèle digne de confiance par design.

Références

[1] B. U. Forstmann, R. Ratcliff, and E.-J. Wagenmakers. Sequential Sampling Models in Cognitive Neuroscience : Advantages, Applications, and Extensions. *Annual Review of Psychology*, 67 :641–666, 2016.

[2] S. Miletić, R. J. Boag, A. C. Trutti, N. Stevenson, B. U. Forstmann, and A. Heathcote. A new model of decision processing in instrumental learning tasks. *eLife*, 10 :e63055, January 2021.

[3] T. J. Pleskac and J. R. Busemeyer. Two-stage dynamic signal detection : a theory of choice, decision time, and confidence. *Psychological Review*, 117(3) :864–901, July 2010.

L'explicabilité appliquée aux modèles de diffusion

Raphael Teitgen, Jeanine Harb, Jeanne Le Peillet

Beink Dream

{raphael.teitgen, jeanine.harb, jeanne.lepeillet}@beink.fr

L'explicabilité est un champ de recherche en intelligence artificielle (IA) [1] développant des outils, techniques et algorithmes visant à fournir des explications claires aux utilisateurs sur les décisions prises par les systèmes d'IA. Cette discipline aspire à rendre les processus décisionnels des modèles d'IA transparents et intelligibles, augmentant ainsi la confiance des utilisateurs et l'efficacité globale de ces technologies. Dans ce contexte, l'AI Act adopté par l'Union européenne régule l'utilisation de l'IA, cherchant à sécuriser les droits des citoyens tout en stimulant l'innovation et l'adhésion aux normes éthiques.

L'explicabilité est particulièrement pertinente pour les modèles d'apprentissage machine, tels que les modèles de diffusion [2], qui synthétisent des images à travers des techniques avancées de traitement de réseaux de neurones. Ces modèles introduisent initialement un bruit dans des images avant de les reconstituer via un processus de "reverse diffusion". Ce processus repose sur des techniques de débruitage qui utilisent des descriptions contextuelles pour guider la reconstitution de l'image. En faisant appel à des embeddings textuels issus de modèles de traitement du langage naturel, le modèle est capable de comprendre et d'interpréter la description textuelle, ce qui influence directement la reconstruction de l'image. Cette capacité de générer ou de modifier des images selon des critères spécifiques rend ces modèles extrêmement utiles pour des applications créatives ou de personnalisation d'images.

Cependant, la nature multimodale de ces entrées, comprenant des descriptions textuelles, des images sources et des masques, ajoute une couche de complexité qui masque les mécanismes internes des modèles, renforçant ainsi leur nature opaque ou de "boîte noire". Le prompt engineering [3], malgré son but d'optimiser les résultats en affinant les entrées, peut introduire une variabilité supplémentaire dans les images générées. En effet, même de légères variations dans les prompts peuvent entraîner des différences notables dans les résultats, complexifiant la prédiction des sorties. De plus, le transfert d'apprentissage, qui vise à améliorer les performances en pré-entraînant les modèles sur de vastes jeux de données avant de les spécialiser via le fine-tuning, peut également obscurcir le fonctionnement du modèle. Les interactions complexes entre les caractéristiques pré-apprises et les ajustements spécifiques lors du fine-tuning ajoutent une dimension supplémentaire d'opacité.

Face à ces défis, des recherches récentes s'attachent à clarifier l'impact des données d'entraînement sur les performances des modèles [5] et à identifier les biais potentiels [6]. Ces efforts incluent l'utilisation d'outils d'interprétation comme SHAP [8], qui permettent d'évaluer comment chaque entrée influence les décisions du modèle. Ces outils, bien que efficace dans certains cas d'usage, doivent être adaptés pour mieux gérer la complexité inhérente aux modèles génératifs, et tout particulièrement aux modèles de diffusion. L'objectif de ces recherches est de développer des méthodes qui non seulement clarifient le processus décisionnel des IA mais aussi ajustent les modèles pour minimiser les erreurs et les biais, augmentant par là même la transparence et la fiabilité de ces modèles.

Références

[1] A. Das, P. Rad (2020). Opportunities and challenges in explainable artificial intelligence (xai) : A survey. arXiv preprint arXiv :2006.11371.

[2] J. Ho, A. Jain, P. Abbeel (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.

[3] S. Witteveen, M. Andrews (2022). Investigating prompt engineering in diffusion models. arXiv preprint arXiv :2211.15462.

[4] Y. Hao, Z. Chi, L. Dong, F. Wei (2024). Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems, 36.

[5] Z. Dai, D. K. Gifford (2023). Training data attribution for diffusion models. arXiv preprint arXiv :2306.02174.

[6] M. Pennisi, G. Bellitto, S. Palazzo, M. Shah, C. Spampinato (2024). Diffexplainer : Towards Cross-modal Global Explanations with Diffusion Models. arXiv preprint arXiv :2404.02618.

[7] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, M. Du. (2024). Explainability for large language models : A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

[8] S. M.M Lundberg, S. I. Lee (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Session 2 : sciences humaines

La normalisation de l'IA : un déluge de réinterprétations de l'AI Act

Hélène Herman[1], Mélanie Gornet[2]

[1] Centre d'Étude des mouvements sociaux, CEMS
[2] Institut Polytechnique de Paris, IP Paris

helene.herman@ehess.fr

Introduction

Il y a d'une part des études notables sur les agents de l'Union Européenne – notamment sur les lobbyistes (Laurens, 2015) ou les experts de la Commission européenne (Robert, 2010) et d'autre part des études sur les processus de transformation de l'intelligence artificielle en problème public (Bellon, 2023) et en objet à légiférer [1]. Le présent article s'attache à analyser les stratégies des normalisateurs de l'intelligence artificielle, dont le travail est touffu, technique, méticuleux. Il est le fruit d'une enquête en cours sur le travail et les membres du groupe de travail « Foundational and societal aspects » de la CEN-CENELEC. Suite à la publication de l'AI Act, une commande a été passée par la Commission Européenne auprès de la CEN-CENELEC pour traduire le règlement en normes para-réglementaires. Celles qui seront « harmonisées » auront valeur de décret d'application de la loi, et vont véritablement structurer le marché de l'intelligence artificielle. La normalisation est souvent méconnue, parce que ces négociations laborieuses n'ont pas d'autre public qu'une majorité d'industriels. L'impact des choix technologiques est souvent sous-estimé, comme ça a déjà été observé sur la gouvernance d'Internet (Lessig, 1999). En effet, il ne s'agit pas d'un processus de traduction comme défini dans la théorie de l'acteur-réseau (Latour, 1984 ; Callon, 1986), mais plutôt de réinterprétation.

Un objet de normalisation particulier

En effet, l'intelligence artificielle n'est pas un pas un objet de normalisation ordinaire. Premièrement, l'aspect définitionnel occupe une bonne partie du travail de délibération : s'entendre sur des notions telles que « l'IA de confiance » n'est pas une mince affaire. Deuxièmement, fixer des seuils durs comme peut d'ordinaire le faire la science réglementaire paraît dérisoire étant donnée la vitesse à laquelle évolue la technologie. Alors se dessinent plutôt des obligations molles de moyens et de protocoles.

Des acteurs privés à la manœuvre

Dans la diplomatie technique qu'est la normalisation, le champ est largement laissé libre aux acteurs privés, qui cherchent à rendre les normes compatibles avec les produits et services qu'ils proposent. Le travail effectué à la CEN-CENELEC est un travail volontaire bénévole sur le temps long, qui suppose des ressources financières et humaines. Ainsi, ce ne sont pas des experts indépendants que l'on trouve à la table de négociation mais des professionnels, qui ont souvent plus d'une affiliation institutionnelle, dont l'adhésion est sponsorisée. Ils peuvent y produire des propositions et doivent aboutir à un consensus. Il s'agit d'une activité hybride entre éléments scientifiques, jugement sociopolitique sur l'acceptabilité du risque (Jasanoff, 1990) et intérêts d'une entreprise.

Se démarquer dans le déluge normatif

On peut constater aujourd'hui une prolifération de normes relatives à l'intelligence artificielle, qui profite aux organismes de normalisation. Ces normes sont en concurrence les unes avec les autres, puisque le système repose beaucoup sur le *self-assessment*, c'est-à-dire que c'est aux entreprises qu'il revient de prouver leur conformité au règlement. C'est tout de même à la Commission Européenne qu'il revient de valider les normes harmonisées, soit celles qui auront valeur de décret d'application de la loi, et vont véritablement structurer le marché de l'intelligence artificielle. Cela participe donc du grand mouvement de régulation, contre-culture de gouvernement, s'appuyant sur un réseau d'autorités de la concurrence [2].

Références

[1] M. Gornet and W. Maxwell. Normes techniques et éthique de l'IA. In *Conférence Nationale en Intelligence Artificielle*, 2023.

[2] A. Vauchez, editor. *Le moment régulateur. Naissance d'une contre-culture de gouvernement*. Presses de Sciences Po., 2024.

IA générative et désinformation : quel impact sur les rapports de force existants en géopolitique ?

Alice Maranne[1*] Clara Fontaine-Say[1] I. Chraibi Kaadoud[2,3]
[1] Indépendante, [2] IMT Atlantique, Brest [3] Centre INRIA de l'Université de Bordeaux

alice.maranne@gmail.com

En 2024, la désinformation et la mésinformation ont été placées comme premier risque mondial par le World Economic Forum devant les risques climatiques[3]. Cela inquiète d'autant plus que 50% de la population mondiale doit se rendre aux urnes durant cette année. Par désinformation, nous entendons l'acte de répandre intentionnellement une information fausse ou manipulée dans le but d'alimenter ou miner une idéologie, concernant des enjeux sociétaux, des débats politiques ou encore des conflits sociaux [1]. Se distinguant de la mésinformation, la désinformation s'inscrit dans une dynamique de guerre de l'information, considérée comme la conduite d'« efforts ciblés » visant à entraver la prise de décision d'un adversaire en portant atteinte à l'information dans son aspect quantitatif (collecte ou entrave à la collecte d'information) aussi bien que qualitatif (propagation ou dégradation) [2]. Dans cette guerre, des rapports de force existent : ils représentent l'équilibre des pouvoirs dans le système international face aux États les plus puissants. Ils peuvent être : (i) interne par le biais de la construction de sa propre force étatique ou (ii) externe avec la recherche d'alliances. Dans nos travaux, nous questionnons l'instrumentalisation de l'IA générative dans les dynamiques de désinformations mondiales et son impact sur les rapports de forces existants ?

Pour ce faire, nous avons réalisé une étude comparative de deux cas d'études, la guerre russo-ukrainienne et le conflit économique autour de Taïwan :

• **Rapports de force** : Une configuration tripartite distingue ces guerres informationnelles : au conflit russo-ukrainien s'ajoute l'OTAN, soutien militaire de l'Ukraine, tandis que pour les tensions sino-taïwanaises, les États-Unis figurent comme allié principal de l'île. Des schémas d'oppositions semblables se distinguent ici : l'Occident contre un régime autoritaire, se disputant un territoire avec un intérêt idéologique ou économique.

• **La quantité et la qualité des narratifs** : Les outils d'IA génératives démocratisent la création de désinformation de meilleure qualité et en grande quantité. Ainsi, lors des élections taïwanaises, 15 000 fausses informations diffusées ont été dénombrées par les experts taïwanais. Dans ce contexte, l'objectif de la Chine a été de mener l'opinion publique taïwanaise vers l'unification voulue par le Parti communiste chinois, au travers de la diffusion de narratifs visant à dépeindre un portrait négatif des États-Unis. De manière similaire en Ukraine, le narratif propagé par la Russie visait à promouvoir l'unicité des deux territoires et des peuples russes et Ukrainiens. Si l'IA générative permet ici d'exacerber des dynamiques de désinformations, ces dernières étaient déjà existantes. Cette technologie ne semble donc pas changer fondamentalement les mécaniques existantes de la désinformation [4].C'est l'alliance de l'IA générative et des plate-formes de diffusion, ou médias alternatifs, qui joue un rôle important dans la propagation rapide et efficace de cette désinformation.

• **La personnalisation et le ciblage** : La diffusion massive de désinformation s'est illustrée également par la multiplicité de *deepfakes* qui est apparue dans les guerres d'informations de toutes natures : pour le conflit ukrainien, on peut citer celui du président Zelensky appelant à déposer les armes et pour Taïwan, celui du candidat élu William Lai qui soutenait la liste d'opposition. Par ailleurs, dans le cas du conflit russo-ukrainien, la diffusion massive de désinformation russe a été personnalisée pour atteindre différents publics en Afrique et au Moyen Orient. Ici c'est l'alliance des *deepfakes* et des technologies de ciblage dans le but d'éroder la confiance dans une institution ou une personnalité politique qui est à relever, en permettant de donner une réalité aux narratifs des stratégies existantes de désinformation.

Notre étude montre que l'IA générative est un nouvel outil au service de la désinformation et non un bouleversement Il faut donc faut penser la désinformation comme un problème géopolitique et non uniquement technologique.

Références

[1] W Bennett and Steven Livingston. *The disinformation age*. Cambridge University Press, 2020.

[2] Dragan Z Damjanović. Types of information warfare and examples of malicious programs of information warfare. *Vojnotehnicki glasnik/Military Technical Courier*, 65(4) :1044–1059, 2017.

[3] The World Economic Forum. Global risks report 2024, 2024.

[4] Felix M Simon, Sacha Altay, and Hugo Mercier. Misinformation reloaded ? fears about the impact of generative ai on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, 4(5), 2023.

*Contact author

L'intelligence artificielle à la lumière de la mythologie grecque : rendre compréhensible les impacts de l'IA pour le grand public

Fabrice Muhlenbach

Université Jean Monnet, UJM-Saint-Etienne, CNRS,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint Etienne, France

fabrice.muhlenbach@univ-st-etienne.fr

Depuis les années 2010, l'intelligence artificielle (IA) est un domaine qui s'est fait connaître du grand public à la faveur du développement des modèles reposant sur le principe de l'apprentissage profond [3]. L'IA ne laisse pas indifférent. Enthousiastes, certains voient en l'intelligence artificielle une fabuleuse histoire composée de toute une série d'inventions témoignant du génie humain [7] alors que d'autres, critiques, la considèrent comme la nouvelle barbarie [1] ou une offensive technologique résolument antihumaniste [8]. En effet, l'utilisabilité de l'IA peut s'avérer être une réelle menace pour les valeurs humaines [6] mais peut aussi apporter une aide précieuse pour trouver des solutions aux grands problèmes que rencontre l'humanité [5].

Face à une mauvaise compréhension de ce que peut faire l'intelligence artificielle, de ses atouts et ses limites, des dangers qui lui sont associés et comment utiliser cette technologie à bon escient, nous proposons une manière originale de présenter l'IA et son fonctionnement sous la forme d'une analogie avec des légendes issues de la mythologie grecque, une source de connaissance plus accessible pour un grand public davantage tourné vers les humanités que vers les sciences du numérique.

Faisant le parallèle entre l'IA et les origines et la destinée du héros mythique Œdipe, nous évoquerons ainsi :

- la fondation mythique de Thèbes par Cadmos et l'introduction de l'alphabet en Grèce, ou comment les performances actuelles de l'IA sont indissociables de la révolution numérique ;
- la trahison de Labdacos, petit-fils de Cadmos, et la malédiction des Labdacides, une famille qui va de travers, tordue comme la lettre λ (lambda/labda) [9], ou l'origine des biais algorithmiques ;
- le crime de Laïos, fils de Labdacos, et son accession au trône de Thèbes, ou l'utilisation de l'IA dans certains contextes malgré les risques encourus ;
- le discours de Laïos rencontrant son fils Œdipe sans le reconnaître, ou les malentendus associés à l'emploi de l'intelligence artificielle générative ;
- Œdipe et l'énigme du Sphinx, ou comment l'IA peut résoudre des problèmes sans les comprendre ;
- Œdipe, parricide et régicide, ou comment l'IA remet en question l'autorité à différents niveaux ;
- Œdipe, époux de la reine Jocaste, sa propre mère, et la rupture de l'interdit de l'inceste [2, 4], ou comment les systèmes automatisés peuvent remettre en question les bases de la civilisation ;
- Œdipe roi, maître de Thèbes, ou comment la société se déleste du poids de la responsabilité et les dangers de la gouvernementalité algorithmique ;
- la peste frappant Thèbes et l'enquête portant sur le meurtre de Laïos, ou la recherche de l'interprétabilité des modèles d'IA ;
- Antigone, fille d'Œdipe, guide de son père aveugle, respectueuse des lois des dieux, ou la réalisation des modèles d'IA éthiques par conception.

Ce tour d'horizon suivant le destin d'Œdipe a comme finalité d'éclairer des non-spécialistes sur les principaux impacts sociétaux liés à l'IA afin de dépasser une vision partisane et que cette technologie soit employée avec raison.

Références

[1] M. David et C. Sauviat. *Intelligence artificielle. La nouvelle barbarie*. Eds. du Rocher, 2019.

[2] S. Freud. *Totem et Tabou*. Points, 2010. Première parution en allemand en 1912.

[3] Y. Le Cun. *Quand la machine apprend*. Odile Jacob, 2019.

[4] C. Lévi-Strauss. *Les structures élémentaires de la parenté*. Mouton & Co, 1947.

[5] F. Mazzi and L. Floridi, editors. *The Ethics of Artificial Intelligence for the Sustainable Development Goals*. Springer, 2023.

[6] C. O'Neil. *Algorithmes : la bombe à retardement*. Les Arènes, 2018. Première parution en anglais en 2016.

[7] C. A. Pickover. *La fabuleuse histoire de l'intelligence artificielle*. Dunod, 2021. 1ère parution en anglais en 2019.

[8] É. Sadin. *L'intelligence artificielle ou l'enjeu du siècle*. L'Échappée, 2018.

[9] J.-P. Vernant. Le tyran boiteux : d'Œdipe à Périandre. In J.-P. Vernant et P. Vidal-Naquet, editor, *Mythe et tragédie en Grèce ancienne – Tome II*, pages 45–72. La Découverte / Poche, 2001. 1ère édition en 1986.

Session 3 : éthique computationnelle

Cadre conceptuel pour les agents autonomes éthiques : application aux agents conversationnels

Robert Voyer[1,2], Thierno Tounkara[1]

[1] LITEM, Univ Evry, IMT-BS, Université Paris-Saclay, 91025, Evry, France
[2] Paris Descartes - LASCO

La définition de l'éthique (individuelle et collective) des agents artificiels nécessite la modélisation des raisonnements éthiques et moraux, au sein des développements informatiques en général et de l'intelligence artificielle plus particulièrement. Pour traiter la problématique de modélisation, nous considérons un cadre conceptuel, appelé « *Modèle des ordres dynamique* », qui propose, une définition structurelle et dynamique de l'éthique individuelle des agents autonomes et peut servir de socle de reflexions, de modélisation ou d'implémentation de l'éthique des agents autonomes ou des agents conversationnels et peut permettre d'enrichir des modèles existants comme le modèle Belief-Desire-Intentions, et aussi des travaux futurs de modélisation de l'éthique collective dans les systèmes multi-agents [3]. La composante structurelle du modèle que nous proposons repose sur le modèle des ordres pascaliens adaptés à notre époque et initialement proposé par le philosophe André Comte-Sponville [4]. Nous avons enrichit le modèle des ordres d'un point de vue logique afin de mieux définir la prise de decision éthique (composante dynamique) du modèle. Nous avons retenu quatre ordres ou niveaux : Science, Politique, Morale et Éthique. Le plus souvent, les ordres, soumis à des principes de structuration internes et indépendants, ne vont pas toujours et partout dans la même direction. Il faut alors choisir, au cas par cas, lorsque les ordres entrent en contradiction, dans telle ou telle situation, quel.s ordre.s choisir en priorité[4]. Les notions de *primat* et de *primauté* vont permettre d'encadrer et de comprendre le choix de l'agent lorsque les ordres se contredisent.Le conflit entre les ordres de la morale et de la politique est le plus fréquent et le plus difficile puisqu'à la différence des autres, il oppose deux ordres de nature axiologique ou normative. Il n'est pas question de valeurs (politiques, morales ou éthiques) qui doivent *nécessairement* prendre appui sur des faits du scientifiquement ou techniquement possible, mais de valeurs politiques et morales qui peuvent entrer en conflit. Le fait scientifique en tant qu'objectivité (axiologiquement neutre) s'impose à la valeur politique ou morale, alors qu'entre deux valeurs (politiques ou morales), non seulement l'accord n'est plus *nécessaire*, mais parfois le conflit révèle une opposition irréductible. Notre étude propose la définition d'un modèle conceptuel UML complet des ordres et de leurs possibles oppositions. Pour illustrer notre démarche, nous appliquons notre modèle aux agents conversationnels. Récemment, avec l'émergence des grands modèles de langage (LLM), une autre approche ascendante de la représentation de la moralité dans les systèmes pourrait consister à former les LLMs sur des textes moralement pertinents, tels que des écrits philosophiques moraux, des fables ou des textes religieux. L'approche de l'IA constitutionnelle, développée et mise en œuvre par [1] est spécifique aux grands modèles de langage (LLM). Dans cette approche, l'apprentissage par renforcement éthique est utilisé pour affiner les réglages des paramètres d'apprentissage, c'est-à-dire après qu'un modèle a été préformé sur un grand nombre de données linguistiques. Selon cette méthodologie, les principes moraux sont explicitement définis par des « promts ». De la même façon, en nous appuyant sur notre modèle cnceptuel, nous avons procédé par la définition précise et détaillée de « prompts » dans les instructions d'un GPT que nous avons spécialement conçu et entraîné avec plusieurs documents décrivant notre cadre conceptuel, des ressources en philosophie morale et des exemples d'application. À noter que notre GPT modèlise non pas un seul agent autonome mais, suivant les orientations envisagés par Open AI [2], orchestre une communauté de trois agents autonomes qui correspondent aux trois ordres : Science, Politique et Morale. Nous appliquons notre GPT à deux situations problématiques différentes ; celle du célèbre dilemme du Trolley ainsi qu'à celle des paradis fiscaux qui sont un exemple de dilemme où la loi autorise une décision alors que la morale l'interdit.

Références

[1] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, and C. McKinnon. Constitutional ai : Harmlessness from ai feedback. *arXiv preprint arXiv :2212.08073*, 2022.

[2] K. C. Sabreena Basheer. Openai's ai agents to automate complex tasks, 2024.

[3] N. Cointe, G. Bonnet, and O. Boissier. De l'intérêt de l'éthique collective pour les systèmes multi-agents. In *Plate-forme intelligence artificielle 2015*, 2015.

[4] A. Comte-Sponville. *Le capitalisme est-il moral ?* Albin Michel, 2004.

Définition de la compatibilité pour des préférences morales : une condition basée sur la cohérence de Suzumura

Guillaume Gervois, Gauvain Bourgne, Marie-Jeanne Lesot

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

prénom.nom@lip6.fr

L'objectif du domaine de l'éthique computationnelle est d'intégrer des principes éthiques au sein des systèmes de prise de décisions. Il peut s'agir de proposer un système qui résout des problèmes éthiques, ou de s'assurer de la conformité d'un système de prise de décisions à des principes éthiques. Ces principes peuvent être définis comme des processus qui, pour un contexte donné, produisent un ordonnancement des décisions réalisables. On obtient alors des préférences morales qui sont formalisées comme des relations binaires sur un ensemble de décisions. Leur partie asymétrique correspond à la comparaison *mieux que* et leur partie symétrique à la comparaison *de valeur égale à*.

Les outils de traitement des préférences peuvent être utilisés pour l'éthique computationnelle (voir par exemple [1]). Une question communément traitée par ces outils est l'agrégation des préférences, qui constitue une problématique que l'on retrouve notamment en choix social computationnel. Il s'agit de chercher un consensus entre les différentes préférences exprimées, quitte à contredire certaines préférences en cas de désaccords.

Néanmoins lorsque l'on considère des préférences morales, deux types de préférences sont à distinguer : celles qui proviennent d'un agent et celles qui proviennent directement des principes formalisés en éthique computationnelle. Les secondes disposent d'un statut particulier car les contredire remet en cause la capacité du principe à fournir des préférences satisfaisantes moralement. C'est pourquoi un désaccord entre différents principes éthiques est plus important qu'un désaccord entre différents agents, et doit faire l'objet d'une attention particulière.

Répondre à la question de leur compatibilité constitue alors une étape préliminaire, qui permet ensuite de déterminer si l'on doit agréger des préférences ou proposer d'autres traitements, comme rejeter un principe ou contraindre ses applications. Cependant, la définition de la compatibilité n'est pas si intuitive car elle dépasse les propriétés usuelles qu'on attend des relations binaires et des préférences.

Dans ce contexte, nous proposons des outils mathématiques pour évaluer la compatibilité d'un ensemble de relations binaires que l'on interprète comme des préférences sur un ensemble de décisions. Cet ensemble constitue notre entrée et nous y faisons référence par le terme de *relations initiales*.

Nous proposons une définition de la compatibilité pour un tel ensemble et nous prouvons des théorèmes qui établissent des conditions à la compatibilité. Nous utilisons pour cela la notion d'extension de relations binaires où une relation x est une extension d'une relation y si les parties symétrique et asymétrique de y sont respectivement incluses dans les parties symétrique et asymétrique de x.

Les relations initiales sont compatibles s'il existe un préordre total, c'est-à-dire une relation transitive et totale, qui soit une extension de toutes les relations initiales. En d'autres termes, elles sont compatibles s'il existe une relation de préférence qui contient simultanément toutes les relations initiales.

Nous montrons que cette définition est équivalente à une version modifiée de la propriété de cohérence proposée par Suzumura [2]. Contrairement à cette dernière, notre version s'applique sur un ensemble de relations binaires. Ce rapprochement fournit une condition vérifiable pour la compatibilité d'un ensemble donné de relations initiales. Nous montrons également que la définition est équivalente à des propriétés faisant intervenir l'union des relations initiales. Ainsi, la compatibilité correspond au cas où l'union est un opérateur d'agrégation satisfaisant. Il s'agit d'un cas rare en choix social computationnel où l'on suppose que des incompatibilités existent parmi les relations de préférences.

Enfin pour un ensemble fini de décisions, nous proposons une preuve alternative à celle proposée par Suzumura en étudiant la construction d'un préordre total qui soit une extension des relations initiales. Nous considérons également l'ajout de contraintes supplémentaires qui imposent des conditions spécifiques que l'extension doit vérifier. Ces contraintes réduisent l'espace des préordres qui sont des extensions totales des relations initiales. Comme pour le cas général, nous proposons une preuve par construction pour deux contraintes simultanément, choisies arbitrairement.

Références

[1] F. Rossi. Moral preferences. In *10th Workshop on Advances in Preference Handling*, 2016.

[2] K. Suzumura. Remarks on the theory of collective choice. *Economica*, 43(172) :381–390, 1976.

Modèle d'éthique pour les MDP multi-agents

Mihail Stojanovski, Nadjet Bourdache, Grégory Bonnet, Abdel-Illah Mouaddib
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
prénom.nom@unicaen.fr

Avec le développement des systèmes d'agents autonomes, des problématiques de décision automatisée peuvent se poser en raison de l'absence de prise en compte d'une composante éthique. De nombreux travaux se sont alors intéressés à intégrer dans les processus décisionnels de ces agents des valeurs morales, des règles morales ou des règles éthiques sur lesquelles ils peuvent fonder leurs décisions. Dans la littérature, plusieurs approches d'éthique computationelle ont été développées, d'un côté des modèles qualitatifs fondés sur du raisonnement logique [1, 2, 4] et de l'autre des modèles quantitatifs fondés sur des processus décisionnels de Markov [3, 5, 6, 8]. Toutefois, les approches quantitatives manquent encore de généricité : les principes éthiques qu'elles mettent en œuvre sont soit représentés par des contraintes *ad hoc* exogènes au modèle de décision lui-même, soit en considérant de manière agrégée ce qui est éthique ou non. Le problème qui en ressort est que le concepteur humain doit utiliser ou créer un modèle différent pour chaque principe éthique, ce qui peut prendre beaucoup de temps et/ou d'efforts et doit représenter de manière implicite les différentes composantes de la prise de décision éthique.

Dans des travaux précédents [7], nous avons proposé un modèle de processus décisionnel markovien éthique (E-MDP) pour un cadre mono-agent. Ce modèle permet d'exprimer différents types de comportements éthiques en s'appuyant sur : (1) un contexte éthique qui indique les valeurs morales positives ou négatives pour l'agent, (2) une évaluation des transitions en termes de promotion ou de violation de chaque valeur morale, (3) une fonction de récompense multicritère qui se fonde sur les deux composants précédents pour représenter explicitement le fait de causer ou réparer du bien ou du mal, (4) un algorithme de résolution lexicographique qui minimise le mal causé par l'agent puis qui cherche à maximiser le bien qu'il peut produire. Cela offre alors un moyen générique de représenter des principes éthiques en se concentrant sur leurs valeurs sous-jacentes, et nous avons proposé des contraintes sur le modèle de transition pour représenter des cadres éthiques comme la théorie du commandement divin, des déontologies fondées sur des devoirs *prima facie* et une éthique de la vertu. Toutefois, ce modèle reste un modèle mono-agent.

Nous proposons ici une extension des E-MDP au cadre multi-agent décentralisé, les E-Dec-MDP, qui ont pour objectif de représenter explicitement la subjectivité éthique de chaque agent et de modéliser la prise en compte de l'impact des décisions individuelles sur la subjectivité des autres. Pour ce faire, chaque agent dispose (1) d'un contexte éthique et (2) d'évaluation des transitions qui lui est propre, mais nous faisons l'hypothèse que cette connaissance est partagée. La fonction de récompense (3) intègre alors trois composantes : le bien et le mal causé ou réparé du point de vue individuel, le bien et le mal causé ou réparé du point de vue des autres agents et un paramètre propre à chaque agent lui permettant de spécifier un compromis entre ses valeurs personnelles et les valeurs des autres agents. Afin de calculer une politique jointe, nous proposons une adaptation de l'algorithme JESP à notre procédure de résolution lexicographique. Nous évaluons ensuite notre approche sur un problème d'allocation séquentielle décentralisée.

Références

[1] F. Berreby, G. Bourgne, and J.-G. Ganascia. A declarative modular framework for representing and applying ethical principles. In *16e AAMAS*, page 96–104, 2017.

[2] N. Cointe, G. Bonnet, and O. Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *15e AAMAS*, pages 1106–1114, 2016.

[3] N. De Moura, R. Chatila, K. Evans, S. Chauvier, and E. Dogan. Ethical decision making for autonomous vehicles. In *IEEE Intelligent Vehicles Symposium*, pages 2006–2013, 2020.

[4] E. Lorini. On the logical foundations of moral agency. In *11e DEON*, volume 7393 of *LNCS*, pages 108–122. Springer-Verlag, 2012.

[5] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. Ethically compliant planning within moral communities. In *4e AIES*, pages 188–198, 2021.

[6] M. Rodriguez-Soto, M. Lopez-Sanchez, and J. A. Rodriguez-Aguilar. A structural solution to sequential moral dilemmas. In *19e AAMAS*, pages 1152–1160, 2020.

[7] M. Stojanovski, N. Bourdache, G. Bonnet, and A.-I. Mouaddib. Processus de décision markoviens éthiques. In *17e JIAF*, pages 177–187, 2023.

[8] J. Svegliato, S. Nashed, and S. Zilberstein. Ethically compliant sequential decision making. In *35e AAAI*, pages 11657–11665, 2021.

Équité subjective par les explications

Sarra Tajouri, Alexis Tsoukiàs

Université Paris-Dauphine - PSL - CNRS, LAMSADE

sarra.tajouri@dauphine.eu, alexis.tsoukias@lamsade.dauphine.fr

1 Cadre décisionnel

Dans ce travail, l'approche proposée s'inscrit dans le cadre de l'équité des processus, en particulier dans des contextes de décisions à fort enjeu tels que l'allocation de ressources financières ou des opportunités économiques.
Dans un processus d'aide à la décision, plusieurs parties prenantes interviennent avec des niveaux de pouvoir et d'agentivité différents. Notre analyse se concentre ainsi sur l'équité envers les individus directement impactés par les décisions prises.

2 Définir l'équité subjective

Sur la base d'une analyse sociologique des limites des mesures statistiques d'équité (non développée ici), nous essayons de redéfinir l'équité avec une dimension subjective qui prend en considération les perspectives des individus.
Une façon de conceptualiser ce changement est de passer d'une approche *top-down* à une approche *bottom-up*. Lorsqu'il s'agit de déterminer ce qui constitue l'égalité, la responsabilité ne devrait pas être confiée exclusivement aux décideurs (ceux qui ont plus de pouvoir dans le processus de décision).
Notre définition de l'équité peut être résumée comme une extension subjective de "equals should be treated equally" de Dwork *et al.* [1] à "individuals perceiving themselves as equals should be treated equally".

Definition 2.1 (Équité subjective individuelle) *Un individu x considère qu'il/elle est traité équitablement si tous les individus qu'il/elle considère comme similaires à lui/elle sont similairement traités.*

$$\phi(x, \psi) \Leftrightarrow \forall y \in S_x, T(M(x, \psi), M(y, \psi)) > \epsilon$$

avec S_x étant l'ensemble des individus que x considère similaires à lui/elle construit selon une mesure de similarité non-symétrique, M étant le mapping entre les individus et leur traitement, T une mesure de similarité entre les traitements.

Par extension, on considère qu'un processus de décision est équitable si toutes les personnes impliquées dans le processus de décision considèrent qu'elles sont traitées de manière équitable.

$$F(X, \psi) \Leftrightarrow \forall x \in X, \phi(x, \psi)$$

3 Explications pour l'équité

L'équité procédurale est fréquemment justifiée par la conformité à un standard normatif. Cette approche adopte une position objective, sous-entendant que l'équité est mesurée une fois et demeure inchangée par la suite. Elle repose sur des règles préétablies supervisées par le décideur.
Dans notre approche, l'équité procedurale subjective passe par l'évaluation, par les individus, de leur situation. Ce sont donc les individus qui accordent la légitimité à un processus jugé comme équitable.
Cela exige de fournir des justifications convaincantes aux utilisateurs, qui peuvent être acceptées ou non. En effet, ces arguments sont réfutables, car ils ne sont pas absolus et peuvent être remis en question, avec l'apport de nouvelles informations ou de changement de perspective notamment. Le niveau d'acceptation d'un argument est subjectif, un même argument peut être persuasif pour certaines personnes mais pas pour d'autres.

Definition 3.1 (Équité subjective par les explications) *Un processus est équitable si les explications sur son déroulement sont convaincantes et acceptées par l'ensemble de la population.*

En se basant sur les travaux d'Habermas sur l'action communicative [2], nous voyons les explications en tant que processus d'interaction sociale entre le décideur et l'utilisateur final. Les modèles d'explication peuvent être :

- stratégique : basé sur des faits fondés objectivement,
- normatif : conformité à une norme, à la loi,
- expressif : présentation subjective de soi-même.

Le modèle communicatif intègre les trois autres modèles que les acteurs peuvent utiliser alternativement pour défendre leurs décisions et faire face aux critiques. L'objectif n'est pas le consensus mais la compréhension et l'acceptabilité.

Références

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, New York, NY, USA, 2012. ACMPress.

[2] J. Harbermas. *The theory of communicative action, volume 2 : Lifeworld and system : A critique of functionalist reason*. Boston, MA : Bacon Press, 1987.