Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur

DOCUMENT D'AUTOÉVALUATION

Équipe ACASA

Campagne d'évaluation 2023-2024 — Vague D

Table des matières

1 INFORMATIONS GÉNÉRALES SUR L'ÉQUIPE ACASA

Nom de l'équipe : Agents Cognitifs et Apprentissage Symbolique Automatique (ACASA)

Responsable de l'équipe : Jean-Gabriel Ganascia, jusqu'au 31 août 2023, puis Gauvain Bourgne

	2017	2018	2019	2020	2021	2022
PR	1	1	1	1	1	1
MCF HDR	1	1	2	2	3	3
MCF	2	2	1	1	0	0
DR	0	0	0	0	0	0
CR HDR	0	0	0	0	0	0
CR	0	0	0	0	0	0
Total permanents	4	4	4	4	4	4
Émérites	0	0	0	0	0	0
Doctorants	2	3	1	1	5	4
Ingénieurs CDD ou hors tutelles	0	0	1	0	0	0
Post-doc, ATER, etc.	6	1	0	0	0	1
Stagiaires	1	2	5	3	1	3
Total non permanents	9	6	7	4	6	8
Total avec émérites	**13**	**10**	**11**	**8**	**10**	**12**
Equivalent temps plein recherche	2.0	2.0	2.0	2.0	2.0	2.0

TABLE 1 – Personnels ACASA sur la période 2017-2022 (au 1er juillet de chaque année)

1.1 Les thématiques scientifiques et leurs enjeux

À l'origine, les thématiques de recherche de l'équipe ACASA étaient centrées sur l'intelligence artificielle, l'apprentissage automatique symbolique, la fouille de données, l'acquisition de connaissances, la découverte scientifique et la créativité. Avec le temps, elles ont évolué et se concentrent maintenant sur trois thématiques : le versant littéraire des humanités numériques, l'*éthique computationnelle* et l'éthique du numérique.

Versant littéraire des humanités numériques

L'activité sur le versant littéraire des Humanités Numériques s'inscrit dans le prolongement des travaux antérieurs de l'équipe ACASA sur la fouille de textes, la classification de textes, l'alignement unilingue, la reconnaissance d'entités nommées et l'extraction de motifs récurrents musicaux et textuels. Cette activité s'est d'abord déroulée au sein du Labex OBVIL, dont Jean-Gabriel Ganascia a été l'un des promoteurs, avant d'en assumer la fonction de directeur adjoint. De nombreuses ressources (doctorants et post-docs) de l'équipe ACASA sont provenues de ce Labex, qui est une action conjointe de l'équipe ACASA du LIP6 et des équipes de recherche en littérature de Sorbonne Université. Il y eut de nombreuses collaborations tant nationales et qu'internationales dans ce contexte, en particulier le projet "Use and Reuse" financé par la Mellon Foundation, dans lequel l'équipe ACASA a collaboré avec le projet *ARTFL* de l'Université de Chicago, le projet ANR Phoebus ou la présence pendant deux ans de Glen Roe, qui a été financé par le prestigieux *Australian Research Council*, avant de devenir professeur dans la faculté des lettres de Sorbonne Université. Ces activités se poursuivent maintenant dans le cadre de l'OBTIC, qui succède au Labex OBVIL au sein de Sorbonne Université.

Dans ce contexte, notre politique scientifique a été de développer une activité durable de développement et de maintenance d'outils pour la communauté scientifique des humanités numériques en collaboration avec différents partenaires, notamment avec le Labex OBVIL puis l'OBTIC, qui lui succède, et avec la BNF avec laquelle tant OBVIL qu'OBTIC collaborent. L'équipe ACASA a aussi été impliquée dans le DIM (Domaine d'Intérêt Majeur) *Sciences du texte et connaissances nouvelles* financé par la région Île-de-France qui visait à promouvoir la recherche sur le versant littéraire des humanités numériques à l'échelle régionale.

Au plan thématique, ces dernières années, les travaux ont plus particulièrement porté sur la détection de motifs stylistiques syntaxiques récurrents [Frontini et al., 2017, Frontini et al., 2018] et de figures sémantiques [Riguet and Mpouli, 2017, Mpouli and Ganascia, 2017], et sur l'étude de l'intertextualité. Inscrits dans la continuité des travaux anciens de l'équipe ACASA sur la détection de citations et de réutilisations [Ganascia et al., 2014], ces recherches se sont poursuivies en collaboration avec l'université de Chicago et ont été étendus à l'étude de graphes de similitudes [Ganascia, 2021].

Dans le passé, des collaborations avec l'Institut des Textes et des Manuscrits Modernes (ITEM-ENS) – laboratoire spécialisé sur la génétique textuelle –, avaient conduit à la réalisation du logiciel MEDITE d'alignement unilingue de

textes. Ce logiciel, toujours utilisé, sert tant aux besoins de la génétique textuelle, pour comparer automatiquement des états de textes, qu'à la mise en œuvre d'éditions électroniques, par exemple de Ramuz ou de Balzac. Une nouvelle collaboration a été engagée avec ce laboratoire dans le cadre du projet Derrida Hexadécimal ; elle porte sur l'exploitation des disques durs du philosophe Jacques Derrida. Il s'agit là d'étendre la génétique textuelle, qui partait des brouillons de papier d'auteurs, aux différentes versions des fichiers électroniques d'un écrivain de l'âge numérique qui rédige directement sur un ordinateur.

Enfin, nous verrons que, dans le but de faire converger les différentes thématiques de l'équipe, nous commençons des travaux sur la détection d'infox, ce qui fait appel à des techniques d'analyse textuelle et de traitement de la langue, dans la continuité de nos travaux sur les humanités numériques, tout en abordant des problématiques éthiques. L'organisation du colloque infox sur Seine en mars 2023 va dans cette direction.

Éthique computationnelle

Le domaine de l'éthique computationnelle vise à la représentation des concepts de la morale et à la simulation de raisonnements éthiques à l'aide d'outils d'IA. La recherche dans ce secteur est assez nouvelle, puisque le besoin de superviseurs éthiques, qui contrôlent les comportements des agents artificiels, est apparu très récemment comme crucial en raison des succès récents des techniques d'IA, en particulier de l'apprentissage automatique supervisé et plus précisément de l'apprentissage profond, et de leur mise en œuvre dans des applications quotidiennes. Insistons sur la notion clé de « superviseur éthique » que nous abordons ici : il s'agit de logiciels qui contraindraient les agents artificiels afin d'assurer qu'ils n'enfreignent pas les règles morales. Cela ne signifie pas que les agents qui en résulteraient seraient, à proprement parler, « éthiques », ce qui relèverait d'un anthropomorphisme outrancier. Il s'agit plus précisément, de faire en sorte qu'un tel superviseur éthique asservisse les comportements des agents artificiels à des considérations morales. Cette recherche est également stimulante d'un point de vue intellectuel, car elle demande de surmonter de nombreuses difficultés qui découlent du caractère conflictuel des normes juridiques et des maximes morales. La modélisation des raisonnements éthiques génère des contradictions logiques entre des devoirs qui, tout en s'opposant, doivent tous être respectés. L'IA basée sur la logique permet, en utilisant des formalismes non monotones comme les logiques des défauts, de surmonter ces contradictions. En outre, les éthiques utilitaristes, très en vogue dans le monde anglo-saxon, évaluent les conséquences d'une action. Dans ce but, il faut prendre en compte les relations de causalité. C'est ce que nous faisons en ayant recourt à des langages d'action. Enfin, l'éthique traite des obligations, des permissions et des interdictions, c'est-à-dire des modalités déontiques, ce qui nécessite l'utilisation de logiques modales spécifiques. Par conséquent, il serait approprié de combiner des formalismes non-monotones, des modèles causaux et des logiques déontiques pour modéliser le raisonnement éthique. C'est ce que nous essayons de faire dans les travaux que nous poursuivons, avec Gauvain Bourgne et les doctorants que nous encadrons sur ce sujet, en particulier Camilo Sarmiento et Yousef Taheri [Taheri et al., 2021].

En outre, Christophe Denis, aborde l'intelligence artificielle explicable (XAI – *explainable AI* en anglais) et argumentative qui jouent l'une et l'autre un rôle si important. En effet, les agents artificiels ne sont pas des entités techniques autonomes et indépendantes. Ce sont des systèmes socio-techniques, c'est-à-dire des dispositifs techniques qui doivent être conçus en regard des organisations sociales dans lesquelles ils s'insèreront. Pour contribuer aux prises de décisions collectives et susciter l'adhésion, il faut que leurs propositions soient justifiées. Autrement dit, il faut relier leurs conclusions aux indices spécifiques qui les motivent et, par là, rendre l'IA explicable.

La première composante du projet porte sur la modélisation du raisonnement éthique avec des techniques d'IA basées sur la logique. La difficulté vient de la triple contrainte de tout raisonnement éthique que nous rappelons ici :

1. pour déterminer la responsabilité d'un agent, on doit prendre en compte les conséquences que l'on peut raisonnablement anticiper de ses actions, ce qui signifie qu'il est nécessaire d'introduire des *modèles causaux* et de les coupler avec des langages d'actions.
2. on doit traiter des règles du devoir, c'est-à-dire de l'obligation, de la permission, de l'omission et de l'interdiction, autrement dit de modalités déontiques. Une façon naturelle d'en tenir compte serait d'utiliser les logiques modales, et plus précisément les *logiques déontiques*. Cependant, les logiques déontiques les plus courantes, avec une sémantique mathématique claire, sont très limitées.
3. il faut surmonter les dilemmes éthiques, c'est-à-dire les conflits de normes, ce qui est très difficile en utilisant les logiques classiques, et a fortiori les logiques déontiques, qui ne parviennent ni les unes, ni les autres, à traiter les incohérences. Nous utilisons des *formalismes non monotones* conçus pour surmonter les contradictions logiques.

Dans le passé, nous avions utilisé un formalisme non-monotone basé sur des modèles stables, à savoir la programmation par ensembles réponses (ASP – *Answer Set Programming*) pour modéliser le raisonnement éthique (cf. [Ganascia, 2007, Ganascia, 2015]) Nous avons aussi eu recourt à des langages d'action (cf. [Berreby et al., 2018]). Cependant, il manque un modèle causal et des modalités déontiques. Certains travaux ont fait appel aux logiques déontiques (par exemple [1]) et même aux logiques déontiques défaisables [4], mais ils ne parviennent guère à gérer la non-monotonie et n'incluent pas de modèles causaux. D'autres ont essayé d'inclure des modèles causaux (cf. [3]), mais ils ne traitent pas vraiment les conflits éthiques et ne font pas usage des modalités déontiques. Notre objectif est de poursuivre ce travail, en particulier celui très préliminaire que nous avions conduit avec la programmation par ensembles réponses (ASP – *Answer Set Programming*) et ce avec des langages d'action pour essayer de surmonter des difficultés mentionnées ci-dessus, à savoir d'être en mesure d'évaluer les conséquences, de traiter les modalités déontiques et de surmonter les conflits de normes.

Nous avons œuvré en ce sens en particulier avec Gauvain Bourgne et Camilo Sarmiento [Sarmiento et al., 2022b] qui ont travaillé sur l'utilisation d'un modèle de causalité couplé à la planification. Les travaux conduits sur la modélisation du raisonnement éthique se poursuivent dans le cadre du projet ANR tri-national RECOMP (Research on Realtime Compliance Mechanism for AI) qui fait collaborer l'équipe ACASA du LIP6 avec l'équipe Japonaise du National Institute of Informatics dirigée par le professeur Ken Satoh et l'équipe Allemande de Institut fur Angewandte Informatik dirigée par le Professeur Adrian Pascke. Ils ont donné lieu à plusieurs publications et plusieurs projets de publications sont en cours.

De plus, on peut noter que notre équipe a acquis une certaine notoriété internationale dans ce secteur assez actif aujourd'hui. À cet égard, on peut mentionner un certain nombre de références à nos travaux. Nous avons aussi organisé un atelier de travail franco-allemand à Paris sur cette thématique en 2022, afin de mettre sur pieds de nouvelles coopérations internationales. À titre d'illustration, la visite, à la rentrée, du Professeur Hannah Ruschemeier, qui viendra passer 6 semaines dans notre équipe atteste de cette reconnaissance. Au reste, le groupe de travail ACE (Approches Computationnelles de l'éthique) qui vient de se constituer dans le GDR RADIA (successeur du GDR IA) et qui est co-animé par Jean-Gabriel Ganascia et Grégory Bonnet, offrira un cadre propice pour la poursuite de ces travaux et pour la mise en place d'un réseau français sur cette thématique.

La deuxième composante de cette thématique consiste à décrire, représenter et organiser les concepts philosophiques utilisés dans le domaine de l'éthique en utilisant une ontologie qui formalise les relations entre ces concepts. L'idée n'est pas seulement de rendre compte des relations logiques entre les concepts comme on le fait habituellement avec une ontologie formelle, mais aussi d'en donner des définitions textuelles, en français et en anglais, et d'associer à chaque concept les références classiques (livres, passages de livres ou articles) des auteurs clés qui mentionnent ces concepts. Nous avons déjà construit une première ontologie de 300 concepts tirés de la littérature en éthique en utilisant un éditeur d'ontologie, à savoir Protégé. Nous avons commencé à annoter ces concepts avec leurs définitions que nous avons extraites de la littérature, par exemple, de la Stanford Encyclopedia of Philosophy, de Wikipedia ou du Dictionnaire d'éthique et de philosophie morale édité par Monique Canto-Sperber.

Conduits par un doctorant en philosophie, Alexandre Bretel, co-encadré par Jean-Gabriel Ganascia de l'équipe ACASA et par un philosophe de l'université de Grenoble, le Professeur Thierry Menissier, les travaux ultérieurs viseront à trouver un accord sur les relations logiques des concepts choisis, leurs différentes définitions et les références auxquelles ils sont associés. Cela se fera en utilisant des techniques de crowdsourcing et de participation qui seront modérées par des philosophes.

Le troisième composante de cet thématique répond au besoin d'ouverture de la « boîte noire » qui consiste d'abord à construire des explications aidant à comprendre les décisions de la machine, puis éventuellement, à les discuter. Ceci est souvent considéré comme l'un des principaux défis éthiques lancés à l'IA, afin que l'utilisateur utilise la machine comme un assistant, tout en conservant la liberté d'accepter ou de refuser ses recommandations, sans les subir aveuglément. Pour ce faire, il a besoin de comprendre les propositions de la machine, c'est-à-dire d'obtenir des justifications de ses conclusions. Ces justifications doivent rendre explicites les liens entre les descripteurs du cas particulier étudié et les conclusions de la machine. Il peut s'agir de montrer à l'utilisateur que si la valeur d'un attribut avait été modifiée, la conclusion aurait également été différente. Il serait également possible de contraindre les procédures d'apprentissage des réseaux de neurones afin de générer des règles symboliques. Il existe bien d'autres façons de rendre les conclusions compréhensibles et dans chaque cas, il existe différentes stratégies de construction d'explications, soit avec un modèle de décision logique appris, soit en extrayant des règles d'un réseau neuronal entraîné. Ce travail se fait en collaboration avec Christophe Denis, maître de conférence à Sorbonne Université et membre de l'équipe ACASA, qui travaille très activement sur ce sujet (cf. [Denis and Varenne, 2022]).

Dans cet ordre d'idée, il semble utile de faciliter les décisions collectives avec différents agents, parmi lesquels il peut y avoir des agents humains et artificiels, en rendant explicites tous les arguments en faveur ou contre

une décision et en rendant les agents capables d'opposer à certains arguments des contre-arguments. Plus généralement, il s'agit de fonder la délibération et la prise de décision collective sur l'argumentation multilatérale. Ce type de recherche exploratoire s'appuie sur des travaux existants développés au LIP6 par Nicolas Maudet et son groupe dans le projet AMANDE financé par l'ANR (cf. [2]). Nous avons déjà eu des échanges entre le projet AMANDE et le projet EthicAA afin d'appliquer cette approche aux délibérations collectives éthiques inspirées des travaux de Jurgen Habermas sur l'éthique de la communication et la démocratie délibérative. Le projet ANR Algojust qui vient d'être accepté devrait offrir un cadre favorable à cette collaboration.

Au reste, notons que Berger Levrault, un éditeur de logiciels et de solutions numériques (cf. https ://www.berger-levrault.com), nous fournit des cas d'usage qui montrent l'applicabilité des approches développées et qui complètent les cas d'usage académiques sur lesquels nous travaillons déjà. Jean-Gabriel Ganascia co-encadre, avec un enseignant de l'Université de Technologie de Troyes, une doctorante, Marion Olivier [Olivier et al., 2022], qui travaille en contrat CIFRE sur l'utilisation de robots de compagnie dans les EHPADs et sur les problématiques éthiques que cela suscite.

Éthique du numérique et de l'intelligence artificielle

Dans la pièce intitulée *R.U.R. - Rossum Universal Robots*, le dramaturge Karel Capek a utilisé pour la première fois le terme « robot » afin de désigner des esclaves mécaniques destinés à faciliter la vie des hommes en réduisant leur charge de travail. Ce drame se termine tragiquement par la soumission de l'humanité à ces travailleurs artificiels, qui gagnent en dignité, par leur travail, ce que les humains perdent, par leur oisiveté. Il en va de même pour la plupart, sinon la totalité, des fictions impliquant des robots : à la fin, les humains sont dominés par les créatures qu'ils ont inventées. Pourquoi ces histoires finissent-elles mal ? Est-ce inévitable ? Et si non, comment éviter une issue fatale ?

Certains avancent que ces histoires finissent mal parce que les machines deviennent conscientes d'elles-mêmes, prennent leur autonomie et se substituent à l'humanité pour dominer la nature. À les examiner de près, de tels scénarios, du moins au regard de l'état d'avancement le la science actuelle, n'ont aucun fondement. C'est ce que Jean-Gabriel Ganascia a essayé de démontrer en se référant aux connaissances scientifiques actuelles en IA et aux perspectives qui s'offrent à tous, dans un livre publié aux éditions du Seuil en 2017 et intitulé *Le mythe de la singularité* [Ganascia, 2017a]. Ce livre a été récompensé par le prix Roberval et a déjà été traduit en coréen, japonais, arabe et portugais. Il a ensuite fait l'objet de très nombreuses publications soit dans des médias à diffusion large, soit dans différentes revues scientifiques du domaine des SHS.

La deuxième hypothèse, plus plausible celle-ci, serait que les machines nous piègent de la même manière que le balai magique a piégé l'apprenti sorcier dans le célèbre poème de Goethe et dans le film Fantasia de Walt Disney. Rappelons très brièvement l'histoire : l'assistant du sorcier utilisait la formule de son maître pour demander à des balais de remplir d'eau une grande piscine, en allant à plusieurs reprises à la fontaine pour remplir des seaux d'eau et en la versant ensuite dans la piscine, car les seaux étaient petits et la piscine était immense, ce qui est à la fois épuisant et ennuyeux. Les balais ont parfaitement exécuté la tâche, mais une fois la piscine remplie, l'apprenti sorcier n'a pas pu se souvenir de la formule magique requise pour arrêter l'action des balais et les empêcher de déverser de l'eau, ce qui fait qu'il créèrent une inondation. Il en va souvent de même avec l'intelligence artificielle : nous l'utilisons pour exécuter certaines tâches, mais parfois, en nous obéissant aveuglément, elle nous conduit à des situations très embarrassantes que nous n'avions pas envisagées.

Pour éviter de tels désagréments, nous devons d'abord nous demander dans quelles situations l'utilisation de l'IA est pertinente et dans quelles autres elle ne l'est pas. En d'autres termes, nous devons nous demander quelles règles de conduite adopter pour l'intelligence artificielle, afin de déterminer les limites de son utilisation. Pour cela, nous devons nous appuyer sur l'éthique, c'est-à-dire sur la discipline philosophique qui réfléchit aux fondements des règles de conduite, et réfléchir à la manière dont elle s'applique aujourd'hui à l'intelligence artificielle. C'est l'objet des travaux que nous menons dans l'équipe ACASA, en collaboration avec des philosophes et des spécialistes de différentes disciplines, notamment des médecins [Denis, 2021, Ganascia, 2019] et des roboticiens, qui appliquent les techniques d'IA. C'est aussi l'objet des travaux que poursuit Jean-Gabriel Ganascia dans différents comités d'éthique institutionnels, dont le CNPEN – Comité National Pilote du Numérique – du CCNE, dont il est membre, le COMETS, qu'il a présidé pendant 5 ans, le comité d'éthique de Pôle emploi et le comité d'éthique de NumAlim – un consortium qui a pour vocation la numérisation des données de l'alimentation – qu'il préside aussi.

Ces travaux sur l'éthique du numérique ont fait l'objet de nombreuses communications soit à destination du grand public, dans les médias ou dans des ouvrages [Ganascia, 2017b, Ganascia, 2022], soit dans des publications plus spécialisées du secteur des Sciences de l'Homme et de la Société, par exemple dans la Revue des Sciences Humaines ou dans le *Oxford Handbook of Ethics of AI*.

2 INTRODUCTION DU PORTFOLIO

- *Élément 1 (publication) :* *Servitudes virtuelles* [Ganascia, 2022], essai philosophique paru aux éditions du Seuil en 2022. Destiné à un public large, l'ouvrage porte sur l'éthique du numérique. Il tente de prendre du recul par rapport aux innombrables chartes d'éthique du numérique et de l'IA parues depuis huit ans. Il a obtenu le prix "recherche universitaire" du livre FIC (Forum International de la Cybersécurité) en 2022.
- *Élément 2 (publication) :* "The Ethics of the Ethics of AI", article de réflexion philosophique sur l'éthique de l'IA co-écrit avec un philosophe américain et publié dans un ouvrage de synthèse important, le "Oxford Handbook of Ethics of AI". L'article est placé au tout début de cet ouvrage de synthèse ; il est d'ailleurs mentionné dans la rubrique "Ethics of Artificial Intelligence and Robotics" de la Stanford Encyclopedia of Philosophy `https://plato.stanford.edu/entries/ethics-ai/`
- *Élément 3 (publication) :* "A Declarative Modular Framework for Representing and Applying Ethical Principles", publié à AAMAS 2017 cet article présente le cadre de raisonnement éthique développé dans l'équipe à l'occasion de la thèse de Fiona Berreby. Il est représentatif des travaux de l'équipe sur la modélisation de raisonnement éthique, posant les bases d'un cadre qui a été développé par nos travaux ultérieurs. Cet article a été depuis cité par un certain nombre d'acteurs de la communauté d'éthique computationnelle, par exemple dans [5].
- *Élément 4 (publication) :* "Action languages based actual causality in ethical decision making contexts.", Il s'agit d'un article publié dans PRIMA 2022, dont une version journal étendue est en cours de soumission. Il propose une sémantique de causalité réelle basée sur un langage d'action. Cet article est représentatif des récents travaux de thèse de Camilo Sarmiento et pose les bases d'un approfondissement du cadre ACE de raisonnement éthique développé dans l'équipe avec une approche plus formelle et expressive.

3 AUTOÉVALUATION DU BILAN

3.1 Autoévaluation de l'équipe

Domaine 2. Attractivité

Référence 1. L'unité est attractive par son rayonnement scientifique et s'insère dans l'espace européen de la recherche.

- Prix Roberval en 2017 pour *Le mythe de la Singularité* de Jean-Gabriel Ganascia
- Prix "recherche universitaire" du livre FIC (Forum International de la Cybersécurité) en 2022 pour *Servitudes virtuelles* de Jean-Gabriel Ganascia.
- participation aux comités de programme des conférences IJCAI (Jean-Gabriel Ganascia et Gauvain Bourgne), AAAI (Gauvain Bourgne), EGC (Jean-Gabriel Ganascia), IC (Jean-Gabriel Ganascia)
- Organisation du *1st Franco-German dialogue on "Perspectives on Ethics of Artificial Intelligence"*, Paris 2-3 juin 2022, Sorbonne Université, Paris. Jean-Gabriel Ganascia
- Organisation WAICOM 2022 : International Workshop on AI compliance mechanism, Sarrebruck, Allemagne déc-22, Gauvain Bourgne
- Invitation d'un membre de l'équipe à l'université de Chicago en juin 2017, puis au mois de mai 2018 et au mois de mai 2019
- Invitation d'un membre de l'équipe à Montréal dans le cadre de l'OBVIA en novembre 2022
- Invitation d'un membre de l'équipe à Séoul, en Corée du 21 au 29 juin 2019.
- un membre de l'équipe est dans de comité de prospective du CIGREF
- un membre de l'équipe a été le représentant officiel de la France lors des 1ère négociation intergouvernementale de la recommandation sur l'éthique de l'IA de l'UNESCO
- un membre de l'équipe a été membre puis président de différents comités d'éthique (COMETS, CNPEN-CCNE, comité d'éthique de pôle emploi...)
- un membre de l'équipe a été membre du comité scientifique du GDR IA
- un membre de l'équipe co-anime avec Grégory Bonnet le GT Approches Computationnelles de l'Éthique du GDR RADIA
- un membre de l'équipe est président de l'AFAS (Association Française pour l'Avancement des Sciences)
- un membre de l'équipe est président du comité d'orientation du CHEC (Cycle des Hautes Études de la Culture)
- un membre de l'équipe co-anime le GT Explicabilite de l'IA du GDR IA

Référence 2. L'unité est attractive par la qualité de sa politique d'accompagnement des personnels.

- Doctorants de l'équipe ACASA : Yousef Taheri (co-encadrement Jean-Gabriel Ganascia, Gauvain Bourgne), Camilo Sarmiento (co-encadrement Jean-Gabriel Ganascia, Gauvain Bourgne),
- Doctorants co-encadrés avec d'autres équipes du LIP6 : Guillaume Gervois (co-encadrement Marie-Jeanne Lesot, Gauvain Bourgne)
- Doctorants d'informatique co-encadrés avec d'autres laboratoires : Marion Olivier (co-encadrement Jean-Gabriel Ganascia, Dimitri Voilmy de l'UTT), Théophile Bayet, (co-encadrement Christophe Denis, Jean-Daniel Zucker), Djes Fresy Bilenga, cotutelle avec Université des Sciences et Techniques de MASUKU (Gabon) (encadrement Christophe Denis), Caouis Kammegne collaboration avec l'Université Cheikh Anta Diop de Dakar (UCAD), Sénégal (encadrement Christophe Denis)
- Doctorants de philosophie : Alexandre Bretel (co-encadrement Jean-Gabriel Ganascia, Thierry Menissier Grenoble), François Levin (co-encadrement Jean-Gabriel Ganascia, Michaël Foessel), Daniele Cavalli (co-encadrement Jean-Gabriel Ganascia, Peter Brugess ENS)
- Professeurs invités : Ken Satoh (juin à décembre 2022), Anders Albrechtslund (en coopération avec l'Institut des Études Avancées - septembre 2022, juin 2023), Thomas Powers (novembre 2018 - janvier 2019)
- Thèses soutenues : Mihnea Tufis, Fiona Berreby, Marine Riguet, Hubert Etienne

Référence 3. L'unité est attractive par la reconnaissance de ses succès à des appels à projets compétitifs.

- Projet ANR tri-national RECOMP – Implication très forte de l'équipe dans le cadre de ce projet tri-partite et tri-national.
- Projet ANR ScientIA
- Participation au DIM (Domaine d'intérêt majeur) *Sciences du texte et connaissances nouvelles* financé par la région Île-de-France
- Participation au Labex OBVIL qui s'est terminé en 2019
- Participation au projet EMOSPACES qui s'est terminé en 2019

Outre les réponses de l'équipe à des appels à projets compétitifs, l'équipe a des contrats avec des industriels pour l'encadrement de doctorants. Ça a été le cas avec Facebook. C'est le cas avec Berger-Levrault.

Domaine 3. Production scientifique

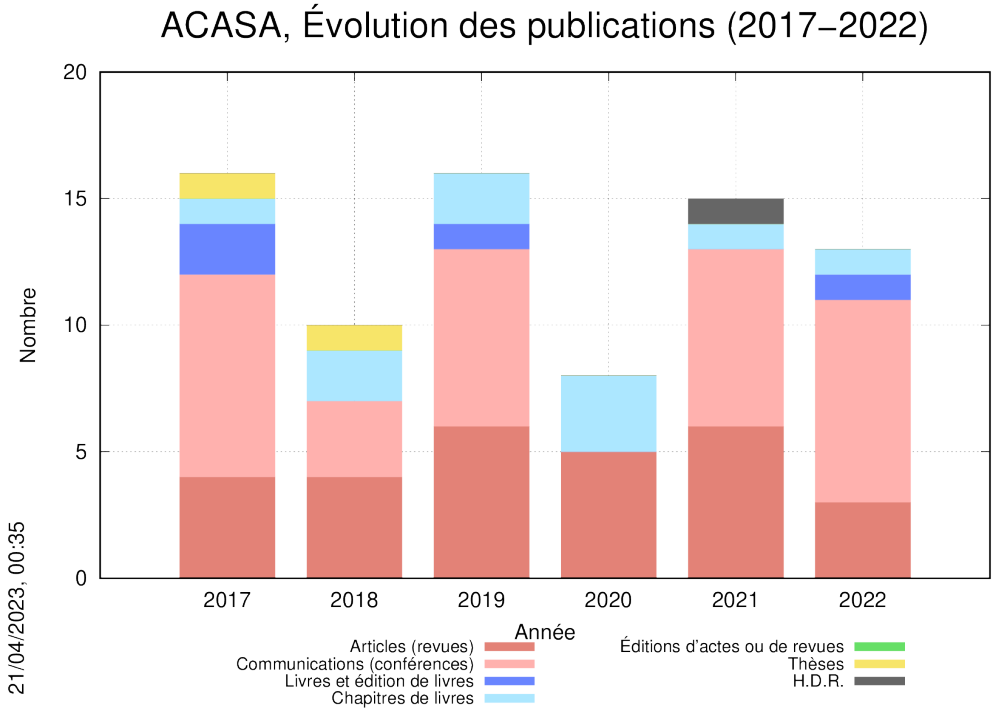

FIGURE 1 – Évolution des publications entre 2017 et 2022

	2017	2018	2019	2020	2021	2022
Articles (revues)	2.00	2.00	3.00	2.50	3.00	1.50
Communications (conférences)	4.00	1.50	3.50	0	3.50	4.00

TABLE 2 – Publications par ETPR par an entre 2017 et 2022

Il faut ajouter la publication de plusieurs essais de réflexion qui ont eu un certain écho dans le public, en particulier "Le mythe de la Singularité" et "Servitudes virtuelles".

Référence 2. La production scientifique de l'unité est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels.

L'équipe privilégie les co-publications avec les doctorants qui participent aux projets de recherche. Dans la mesure du possible, ce sont les doctorants qui vont présenter leurs résultats dans les conférences nationales et internationales.

Domaine 4. Inscription des activités de recherche dans la société

Référence 1. L'unité se distingue par la qualité et la quantité de ses interactions avec le monde non-académique.

L'équipe ACASA a bénéficié régulièrement de bourses CIFRE. C'est aujourd'hui le cas de Marion Olivier qui est financée sur un contrat CIFRE avec la société Berger-Levrault. Ça a été le cas avec Hubert Etienne qui a bénéficié d'un financement CIFRE de la société Facebook/Meta jusqu'à sa soutenance de thèse, en septembre 2022. Il est maintenant employé par la société Meta.

Les membres de l'équipe ont fait de nombreuses interventions sur l'intelligence artificielle et l'éthique de l'intelligence artificielle dans différents contextes professionnels, par exemple, en conférence plénière pour les journées françaises de dermatologie en 2019, ou, toujours en conférence plénière, pour la Société Française de Chirurgie Orthopédique et Traumatologique (SOFCOT) en 2020. Ils interviennent régulièrement pour l'IHEMI (Institut des Hautes Études du Ministère de l'Intérieur) et pour l'IHEDN (Institut des Hautes Études de la Défense Nationale). Un membre de l'équipe est d'ailleurs membre du conseil scientifique de l'IHEMI (institut des hautes études du ministère de l'intérieur).

Référence 2. L'unité développe des produits à destination du monde culturel, économique et social.

Les membres de l'équipe sont impliqués dans différentes institutions : présidence du comité d'éthique de pôle emploi et présidence du comité d'éthique de NumAlim, un consortium qui a pour vocation la numérisation des données de l'alimentation. Présidence du comité d'orientation du CHEC (Cycle des Hautes Études de la Culture).

Référence 3. L'unité partage ses connaissances avec le grand public et intervient dans des débats de société.

L'équipe contribue grandement à la diffusion des résultats de la science dans la société par les ouvrages de ses membres [Ganascia, 2017b, Ganascia, 2017a, Ganascia, 2022], par les articles qu'ils rédigent, par exemple par les chroniques éthiques qui paraissent dans le magazine La Recherche et dans le magazine Sciences et avenir, par les articles qu'ils écrivent pour des journaux à diffusion plus ou moins large, par les interviews auxquels ils répondent assez fréquemment, par les émissions de radio et de télévision auxquelles ils participent. À cela il faut ajouter qu'un membre de l'équipe préside l'Association Française pour l'Avancement des Sciences (AFAS) qui a pour vocation la diffusion scientifique en direction de la société.

4 RÉFÉRENCES BIBLIOGRAPHIQUES EXTERNES

[1] K. Arkoudas, S. Bringsjord, and P. Bello. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pages 17–23, 2005.

[2] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. In *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 1537–1543, Macao, Macau SAR China, August 2019.

[3] J. Y. Halpern and M. Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 1853–1860. AAAI, 2018.

[4] J. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23 :35–65, 1994.

[5] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics : A survey, 2020.

5 RÉFÉRENCES BIBLIOGRAPHIQUES SIGNIFICATIVES DE ACASA

[Berreby et al., 2018] Berreby, F., Bourgne, G., and Ganascia, J.-G. (2018). Event-Based and Scenario-Based Causality for Computational Ethics. In *AAMAS 2018 - 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 147–155, Stockholm, Sweden. International Foundation for Autonomous Agents and Multiagent Systems.

[Denis, 2021] Denis, C. (2021). Le périple de l'éthique de l'Intelligence Artificielle dans la révolution en cours des systèmes de soins. *Journal de Médecine Légale - Droit, Santé et Société*, 3(3) :17–21. https ://www.cairn.info/revue-droit-sante-et-societe-2021-3-page-17.htm.

[Denis and Varenne, 2022] Denis, C. and Varenne, F. (2022). Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine. *Revue Ouverte d'Intelligence Artificielle*, 3(3-4) :287–310.

[Frontini et al., 2017] Frontini, F., Boukhaled, M. A., and Ganascia, J.-G. (2017). Mining for characterising patterns in literature using correspondence analysis : An experiment on french novels. *Digital Humanities Quarterly*, 11(2). `http://www.digitalhumanities.org/dhq/vol/11/2/000295/000295.html`.

[Frontini et al., 2018] Frontini, F., Boukhaled, M. A., and Ganascia, J. G. (2018). Approaching french theatrical characters by syntactical analysis : A study with motifs and correspondence analysis. In Legallois, D., Charnois, T., and Larjavaara, M., editors, *Grammar of Genres and Styles. From Discrete to Non-Discrete Units*, volume 320 of *Trends in Linguistics*, pages 118–139. De Gruyter Mouton, Berlin/Boston.

[Ganascia et al., 2014] Ganascia, J., Glaudes, P., and Lungo, A. D. (2014). Automatic detection of reuses and citations in literary texts. *LLC*, 29(3) :412–421.

[Ganascia, 2007] Ganascia, J.-G. (2007). Modeling Ethical Rules of Lying with Answer Set Programming. *Ethics and Information Technology*, 9(1) :39–47.

[Ganascia, 2015] Ganascia, J.-G. (2015). Non-monotonic Resolution of Conflicts for Ethical Reasoning. In Trappl, R., editor, *A Construction Manual for Robots' Ethical Systems : Requirements, Methods, Implementations*, Cognitive Technologies, pages 101–118. Springer International Publishing, Cham.

[Ganascia, 2017a] Ganascia, J.-G. (2017a). *Le mythe de la Singularité : faut-il craindre l'intelligence artificielle ?*, volume Collection sciences ouvertes. Éditions du Seuil.

[Ganascia, 2017b] Ganascia, J.-G. (2017b). *L'intelligence artificielle : vers une domination programmée ?* collection idées reçues. Le cavalier bleu.

[Ganascia, 2019] Ganascia, J.-G. (2019). L'IA en médecine : oracle, instrument ou ersatz ? *Annales de Dermatologie et de Vénéréologie*, 146(12) :A12–A13.

[Ganascia, 2021] Ganascia, J.-G. (2021). Détection automatique de phénomènes intertextuels. *Genesis (Manuscrits - Recherche - Invention)*, (51) :63–77.

[Ganascia, 2022] Ganascia, J.-G. (2022). *Servitudes virtuelles*. Seuil.

[Mpouli and Ganascia, 2017] Mpouli, S. and Ganascia, J.-G. (2017). Another Facet of Literary Similes : A Study of Noun+Colour Term Adjectives. *CORELA - COgnition, REprésentation, LAngage*, (HS-21).

[Olivier et al., 2022] Olivier, M., Rey, S., Voilmy, D., and Ganascia, J.-G. (2022). Contributions of user tests in a Living Lab in the co-design process of human robot interaction. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Naples, France.

[Riguet and Mpouli, 2017] Riguet, M. and Mpouli, S. (2017). At the crossroads between the scientific and the literary discourse : Comparison as a figure of dialogism. *Digital Scholarship in the Humanities*, 32(supp 2) :ii60–ii77.

[Sarmiento et al., 2022b] Sarmiento, C., Bourgne, G., Inoue, K., and Ganascia, J.-G. (2022b). Action Languages Based Actual Causality in Decision Making Contexts. In *The 24th International Conference on Principles and Practice of Multi-Agent Systems*, volume 13753 of *Lecture Notes in Computer Science*, pages 243–259, Valence, Spain. Springer International Publishing.

[Taheri et al., 2021] Taheri, Y., Bourgne, G., and Ganascia, J.-G. (2021). A Compliance Mechanism for Planning in Privacy Domain Using Policies. In *Fifteenth International Workshop on Juris-informatics (JURISIN 2021)*, Kanagawa, Japan.

A ANNEXE — MEMBRES PERMANENTS AU 31/12/2022

La table ci dessous liste les membres permanents de l'équipe ACASA.

NOM	Prénom	Corps	Employeur
BOURGNE	Gauvain	MCF (HDR)	Sorbonne Université
DENIS	Christophe	MCF (HDR)	Sorbonne Université
FAUCHER	Colette	MCF (HDR)	Sorbonne Université
GANASCIA	Jean-Gabriel	PR	Sorbonne Université

ÉLÉMENT DE PORTFOLIO 01

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : L'ouvrage "Servitudes virtuelles" [1]

URL de l'élément : `https://nuage.lip6.fr/s/zjeBYTGPoBFz5ET`

Fichier de élément : dernières épreuves de l'ouvrage

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un essai philosophique paru aux éditions du Seuil en 2022. Destiné à un public large, l'ouvrage porte sur l'éthique du numérique. Il tente de prendre du recul par rapport aux innombrables chartes d'éthique du numérique et de l'IA parues depuis huit ans. Il a obtenu le prix « recherche universitaire » du livre FIC (Forum International de la Cybersécurité) en 2022.

3 PRÉSENTATION DE CET ÉLÉMENT

À l'évidence, une réflexion sur les conséquences sociales et politiques des technologies de l'information et de la communication et sur les moyens de se délivrer des nouvelles formes d'oppression qu'elles génèrent s'impose. Beaucoup aujourd'hui en sont convaincus : de grands acteurs de l'internet, des sociétés savantes, des universités, des États, des organisations non gouvernementales telle l'UNESCO, des associations, des pouvoirs supranationaux comme la communauté et le parlement européen réunissent des comités d'experts qui rédigent des rapports, édictent des chartes, promulguent des codes et votent des résolutions. Un article paru en 2022 [1] comptabilisait 351 rapports sur l'éthique du numériques et pas moins de 521 recommandations !

Ce faisant, ces groupes invoquent des principes, des droits fondamentaux et des idéaux comme l'autonomie ou la dignité de la personne. Il s'ensuit un certain nombre de recommandations irréprochables contre la déshumanisation, l'émancipation ou la malfaisance des machines. De prime abord, les principes retenus, les droits humains invoqués, les notions mentionnées et les recommandations formulées paraissent louables. Comment s'opposer à la dignité et à l'autonomie de la personne ? Comment ne pas condamner la malfaisance et ne pas aspirer aux bénéfices que nous pourrions tirer de l'usage des machines ? Comment ne pas de réjouir des opportunités offertes par l'intelligence artificielle, tout en prenant garde aux mauvaises utilisations qui peuvent en être faites ? Comment enfin ne pas louer l'idéal de transparence ? Pourtant, l'examen de situations concrètes montre que ces principes, ces droits, ces concepts et ces recommandations demeurent bien abstraits aujourd'hui. Ils renvoient à des craintes ancestrales qui relèvent plus de mythologies et de fables que de réalités effectives ou prévisibles. Dans le même temps, ils nous laissent démunis dans le monde qui se fait jour sous nos yeux, avec le développement de l'informatique et de l'intelligence artificielle, et impuissants face aux nouvelles vulnérabilités qu'il engendre. Ce livre tente de saisir le décalage entre ces idéaux momifiés et la réalité.

À cet effet, la première partie du livre parcourt ce monde nouveau et en décrit les tendances marquantes. Pour ce tour d'horizon, il recourt à une « rose des vents numériques » construite à partir des deux lignes de fuites, la connectivité et la vie. Il s'agit alors d'explorer les quatre points cardinaux induits par ces deux directions : le « hors vie » qui répond aux tentatives de réparation et de remplacement du vivant par le numérique ; le « en ligne » qui vise à connecter chacun d'entre nous à tous les autres, sans répit ni intériorité possible ; le « en vie » qui transforme le tissu social en réinventant — l'ouvrage parle ici de réontologisation — les notions qui en font la trame comme l'amitié, la réputation, la confiance, l'argent ou le travail ; enfin, le « hors ligne » qui se présente parfois à nous comme rêve illusoire d'un échappatoire possible à l'emprise du numérique, mais il constitue bien plus un exil auquel la plupart se sentent condamnés, par devers eux.

Dans la seconde partie, l'ouvrage se propose d'examiner quelques-unes des idées fétiches invoquées au nom de l'éthique. Il s'agit alors de s'interroger sur le sens des règlementations qui, au nom de l'éthique, imposent

1. `https://www.seuil.com/ouvrage/servitudes-virtuelles-jean-gabriel-ganascia/9782021440317`

d'innombrables règlements faits de contraintes juridiques, de normes et de standards. Plus précisément, on y aborde, tour à tour, chacun des quatre idéaux qui servent de fondement à la plupart des chartes d'éthique de l'intelligence artificielle et du numérique et qui s'inspirent de la bioéthique, à savoir l'autonomie, le bienfaisance et la non-malfaisance, la justice et enfin la transparence. On essaie alors de montrer qu'en dépit de l'utilisation qui en est faite, ils ne permettent pas de fonder une éthique du numérique.

Enfin, la conclusion met en regard les nouvelles servitudes consécutives aux développements du numérique avec les servitudes anciennes, et suggère des pistes pour s'en délivrer.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Lionel Nganyewou Tidjon and Foutse Khomh. The different faces of ai ethics across the world : A principle-implementation gap analysis, 2022.

ÉLÉMENT DE PORTFOLIO 02

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : The Ethics of the Ethics of AI

URL de élément : `https://nuage.lip6.fr/s/dBcE87XBkFnLBcx`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un article de réflexion philosophique sur l'éthique de l'IA co-écrit avec un philosophe américain et publié dans un ouvrage de synthèse important, le "Oxford Handbook of Ethics of AI". L'article est placé au tout début de cet ouvrage de synthèse ; il est d'ailleurs mentionné dans la rubrique "Ethics of Artificial Intelligence and Robotics" de la Stanford Encyclopedia of Philosophy `https://plato.stanford.edu/entries/ethics-ai/`.

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article de philosophie aborde plusieurs questions d'éthique qui devraient être préalables à toute éthique de l'IA. Ces questions se répartissent en cinq grandes catégories. La première porte sur les ambiguïtés conceptuelles des notions fondamentales de l'IA lorsqu'elles sont employées par des philosophes ou des juristes. Elles sont dues, en partie, à ce que des termes sont utilisés indifféremment en philosophie et en IA alors qu'ils ont des sens différents dans les différentes communautés. Ainsi en va-t-il par exemple, de la notion d'"agent" ou de la notion d'"autonomie".

La deuxième catégorie de questions est relative à l'estimation des risques, parfois surévalués, parfois sous-estimés. Cette question se révèle d'autant plus cruciale que l'IA-act européen se fonde sur l'anticipation de risques.

La troisième porte sur la mise en œuvre de superviseurs éthiques dans un contexte opérationnel.

La quatrième est relative à l'ambivalence de l'idée d'explication dans le cadre des systèmes d'apprentissage machine entraînés avec de très grandes masses de données. En effet, si l'IA dite explicable est très à la mode, on confond trop souvent la transparence, qui doit rendre compte fidèlement du fonctionnement d'un système, sur chaque cas particulier, au risque d'être opaque, et l'interprétation qui s'éloigne du fonctionnement effectif.

Enfin, la cinquième catégorie de question porte sur l'opposition entre deux façons de voir l'IA : une vue oppositionnelle, où l'on suppose que la machine est rivale, et une vue coopérative, où elle est un partenaire.

Le chapitre montre ensuite que beaucoup d'approches de l'éthique de l'IA n'aborde pas clairement, et de façon argumentée, ces différentes questions. Et, en conséquence, qu'il faudrait renouveler les approches de l'éthique de l'IA. Enfin, l'article aborde les difficultés à surmonter pour réaliser des superviseurs éthiques susceptibles d'imposer des prescriptions aux actions des machines.

ÉLÉMENT DE PORTFOLIO 03

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : A Declarative Modular Framework for Representing and Applying Ethical Principles

URL de l'élément : `https://hal.sorbonne-universite.fr/hal-01564675`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un article publié à AAMAS 2017 présentant le cadre de raisonnement éthique développé dans l'équipe à l'occasion de la thèse de Fiona Berreby. Il est représentatif des travaux de l'équipe sur la modélisation de raisonnement éthique, posant les bases d'un cadre qui a été développé par nos travaux ultérieurs. Cet article a été depuis cité par un certain nombre d'acteurs de la communauté d'éthique computationnelle (une trentaine de citations externes, dont une dizaine données en référence plus bas).

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article examine l'utilisation de langages de haut niveau dans la conception d'agents autonomes éthiques. Il propose un cadre logique nouveau et modulaire pour représenter et raisonner sur une variété de théories éthiques, sur la base d'une version modifiée du calcul des évènements, implémentée en Answer Set Programming.

Le processus de prise de décision éthique est conçu comme une procédure en plusieurs étapes, identifiant les différents composants qu'il est nécessaire de représenter pour permettre un raisonnement complet.

Un modèle d'action, fondé sur des mécanismes classiques de représentation de l'action et du changement, permet à l'agent d'évaluer son environnement et d'anticiper le déroulé des évènements résultant de ses choix, obtenant ainsi pour chaque scénario envisagé (choix possible d'actions par l'agent) une trace des évènements qui se déclenchent dans le système et des états résultants.

Un modèle de causalité permet ensuite d'analyser cette trace pour identifier les relations causales entre les actions de l'agents et les différents évènements pouvant survenir, lui permettant de raisonner sur sa responsabilité, d'identifier conséquences de ses actes et d'estimer si certains effets sont utilisés comme moyens d'arriver à d'autres.

Enfin, le modèle éthique à proprement parler, séparé en modèles du Bien et modèles du Juste déterminent à partir des informations précédentes quels sont les choix éthiquement acceptables selon différents principes éthiques. L'article en présente un certain nombre, tirés de la littérature en éthique normative, intégrant dans un même cadre des principes conséquentialistes et d'autres plus déontologiques.

L'ambition de cette approche est double. Tout d'abord, elle est de permettre la représentation systématique d'un nombre illimité de processus de raisonnements éthiques, à travers un cadre adaptable et extensible. Deuxièmement, elle est d'éviter l'écueil trop courant d'intégrer directement l'information morale dans de raisonnement général sans l'expliciter, alimentant ainsi les agents avec des réponses atomiques qui ne représentent pas la dynamique sous-jacente. En séparant clairement les différentes problématiques de représentation (action, causalité, éthique) et en identifiant explicitement à chaque étapes les entrées du modèle spécifiques à un domaine donné, nous visons à déplacer de manière globale le processus de raisonnement moral du programmeur vers le programme lui-même.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Jugement éthique dans le processus de décision d'un agent bdi. *Rev. d'Intelligence Artif.*, 31(4) :471–499, 2017.

[2] Louise A Dennis, Martin Mose Bentzen, Felix Lindner, and Michael Fisher. Verifiable machine ethics in changing contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11470–11478, 2021.

[3] Louise A Dennis and Cristina Perea del Olmo. A defeasible logic implementation of ethical reasoning. In *First International Workshop on Computational Machine Ethics (CME-2021)*, 2021.

[4] Abeer Dyoub, Stefania Costantini, Francesca Alessandra Lisi, and Ivan Letteri. Logic-based machine learning for transparent ethical agents. In *CILC*, pages 169–183, 2020.

[5] David Fuenmayor and Christoph Benzmüller. Normative reasoning with expressive logic combinations. In *ECAI 2020*, pages 2903–2904. IOS Press, 2020.

[6] Umberto Grandi, Emiliano Lorini, Timothy Parker, and Rachid Alami. Logic-based ethical planning. *arXiv preprint arXiv :2206.00595*, 2022.

[7] Martin Jedwabny, Pierre Bisquert, and Madalina Croitoru. Generating preferred plans with ethical features. In *Florida Artificial Intelligence Research Society*, volume 34, 2021.

[8] Emiliano Lorini. A logic of evaluation. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 827–835, 2021.

[9] Sylvie Michel, Sylvie Gerbaix, and Marc Bidan. A plea for choosing ex ante an ethical theorical position for a relevant response to ethical issues posed by algorithmic systems. In *2022 3rd International Conference on Next Generation Computing Applications (NextComp)*, pages 1–6. IEEE, 2022.

[10] Emery A Neufeld, Ezio Bartocci, Agata Ciabattoni, and Guido Governatori. Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4) :43, 2022.

[11] Maurice Pagnucco, David Rajaratnam, Raynaldio Limarga, Abhaya Nayak, and Yang Song. Epistemic reasoning for machine ethics with situation calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 814–821, 2021.

[12] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics : A survey. *ACM Computing Surveys (CSUR)*, 53(6) :1–38, 2020.

[13] John Zoshak and Kristin Dew. Beyond kant and bentham : How ethical theories are being used in artificial moral agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

ÉLÉMENT DE PORTFOLIO 04

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Action languages based actual causality in ethical decision making contexts

URL de l'élément : `https://hal.science/hal-03790699v1`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un article publié dans PRIMA 2022, dont une version journal étendue est en cours de soumission. Il propose une sémantique de causalité réelle basée sur un langage d'action. Cet article est représentatif des récents travaux de thèse de Camilo Sarmiento et pose les bases d'un approfondissement du cadre ACE de raisonnement éthique développé dans l'équipe avec une approche plus formelle et expressive.

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article propose un langage d'action expressif permettant de modéliser des préconditions riches, des actions concurrentes et le déclenchement d'évènements exogènes tout en conservant des évènements déterministes et instantanés. Il peut être vu un comme un fragment de PDDL+ évitant la complexité des actions duratives. Il permet de déterminer l'évolution du monde étant donné un ensemble d'actions choisies délibérément par des agents et dont l'occurrence peut entraîner une réaction en chaîne d'évènements dit exogènes. Cette possibilité d'avoir des réactions en chaîne, allié à la concurrence des actions, permet de représenter finement des situations causales complexes.

Sur la base de ce langage d'action, l'article présente une définition sémantique formelle de causalité réelle basée sur la notion de NESS cause directe (Necessary Element of a Sufficient Set) et discute du bien-fondé de cette approche par rapport aux classiques approches contrefactuelles en montrant que le fait de se baser sur un langage d'action où les effets à court terme fondent la causalité, il est possible de réfuter les principales critiques apportées aux approches causales par régularité. La NESS cause directe définie entre des évènements et une formule d'état (formule basé sur des fluents) est ensuite élargie à une notion de NESS cause (non forcément directe) permettant d'aller au-délà des évènements déclencheurs directs pour remonter aux actions ayant contribué à la réalisation de la formule. Cette notion est enfin utilisée pour définir une notion de causalité réelle entre deux évènements, en se fondant sur les conditions de déclenchement de l'évènement causé.

La version étendue de cet article propose de plus une traduction correcte et complète de cette sémantique en ASP, permettant le calcul de toutes les relations causales existantes entre les différents évènements d'un scénario donné, permettant ainsi de remplacer le modèle causal initial du cadre de raisonnement éthique ACE développé dans l'équipe (auparavant limité à des théories d'actions exprimées en STRIPS) par ce mécanisme plus fin et expressif.