Rapport de mi-stage

Ethical issues in multi-objective reinforcement learning

Stagiaire :
Marius Le Chapelier

Encadrants :
Aurélie Beynier
Paolo Viappiani
Nicolas Maudet

May 17, 2022

1 Modalités du stage

Le stage a débuté le 7 mars et se termine le 5 septembre, il se déroule au sein de l'équipe Système Multi-Agents (SMA) du laboratoire du Lip6, à Sorbonne Université.

2 Descriptif du sujet

L'IA est de plus en plus présente dans notre société et l'accroissement de son utilisation dans les domaines sensibles (justice, recrutement, voitures autonomes, prêts de banques, etc) et l'augmentation des interactions avec les humains ont décuplé l'intérêt porté à l'integration des problématiques éthiques dans l'IA. En effet, apprendre à des agents à agir, selon les normes morales de la société, n'est pas aisé, car ces normes ne sont pas explicites. Un premier travail serait donc d'estimer ces normes, mais comme aucune estimation n'est assurée d'être parfaite, créer un modèle qui apprend à un agent à suivre les normes sociales sans erreur est complexe et c'est justement l'objet de ce stage.

3 Objectifs du stage

L'objectif pour la fin du stage est de développer une approche d'apprentissage par renforcement multi-objectif respectant un cadre éthique. Pour developper cette approche, il faudra étudier comment différents objectifs éthiques peuvent être appris et combinés pour créer des comportements éthiques et interprétables.

Pour arriver à cette finalité, le stage va se dérouler en plusieurs phases :

D'abord, la réalisation d'un état de l'art de tous les domaines directement en rapport avec le sujet : L'intégration de d'éthique dans l'IA, L'apprentissage par renforcement appliqué à un cadre éthique, l'apprentissage par renforcement

multi-objectif, l'explicabilité des systèmes de décisions. D'autres domaines ne sont pas directement en lien avec le sujet, mais il sera développé par la suite pourquoi il était intéressant de les inclure dans l'état de l'art : l'aprentissage par préferences, l'apprentissage par renforcement inversé, Les generaive adversarial networks (GANs), l'apprentissage par renforcement profond.

Ensuite, l'identification des contributions possibles vis-à-vis des modèles d'apprentissage multi-objectifs sur l'éthique des actions déjà existants.

Dans un troisième temps, il faudra développer un nouveau modèle répondant à certaines contributions identifiées précédemment.

Enfin, il faudra expérimenter et ajuster le modèle, notamment en testant plusieurs techniques d'optimisation. Cette phase d'experimentation devra mettre l'accent sur l'explicabilité des résultats, et particulièrement des comportements appris.

4 Thématiques du stage

4.1 Comment intégrer de l'éthique dans l'IA ?

Les systèmes de décisions basés sur l'IA font de plus en plus partie de nos vies, les laisser prendre des décisions importantes (justice, recrutement, voitures autonomes, prêts de banques, etc) à notre place va passer à un moment ou un autre par l'intégration de problématiques éthiques dans les domaines de l'IA.

La question de l'intégration de l'IA dans l'éthique est complexe, car il n'y a pas de définition objective de ce qu'est l'éthique : il existe différentes définitions philosophiques (déontologie, utilitarisme, etc). Comme ces différentes définitions peuvent être contradictoires ou incomparables (au sens d'une agrégation mathématique [16]), on peut se poser la question de comment, dans un MDP par exemple, valuer la composante éthique d'une action. La réponse n'est pas triviale et plusieurs approches ont été utilisées, chacune possédant ses avantages et inconvénients.

Les approches dites **rule-based ou top-down** [8, 21, 26, 29, 25] :
On inscrit les règles ou contraintes sociales, à priori, lors de la création du modèle, et l'apprentissage se fait dans le cadre de ces règles éthiques.
Avantages : Assurance du respect des contraintes éthiques et explicabilité du comportement.
Inconvénients : Flexibilité des contraintes, énumération de toutes les contraintes à priori (potentiellement impossible), insensibilité aux différences de définitions/valeurs éthiques selon la culture, le contexte ou même le temps (le modèle reste le même et ne peut pas évoluer de lui-même).

Les approches dites **data-driven ou bottom-up** [14, 17, 22, 30, 27] :
On considère que les humains agissent en moyenne de manière éthique et donc que leurs choix peuvent être la base d'un apprentissage supervisé.
Avantages : Le modèle dépend des données en entrée et est donc flexible et adaptable selon les environnements, contextes, cultures, il peut également s'adapter à l'évolution des moeurs de la société en continuant son apprentissage durant son déploiement.
Inconvénients : Les résultats du modèle dépend grandement de la qualité des données en entrée, ce qui induit une difficulté à expliquer le comportement

général, et à assurer que l'agent respecte effectivement les valeurs éthiques. Un autre inconvénient est que pour avoir des résultats réalistes, il faut une grande quantité données réelles d'humains, ce qui pourrait passer par des questionnaires ou autres systèmes de collecte de données (Moral Machine, etc), ou raisonner sur des données simulées lors de la création du modèle (ce qui est fait la plupart du temps).

Les systèmes et approches divergent car trouver un système de valuation éthique des actions est compliqué, mais pour commencer, est-ce la bonne approche ? Est-ce une bonne idée d'essayer de donner une valeur éthique objective à des actions (ce qui induit un ordre complet sur un ensemble d'actions) ? Valuer de manière objective une action, dans le cadre d'un apprentissage automatique, n'irait-il pas à l'encontre même des valeurs éthiques sociales ? Toutes ces questions sont posées dans l'article [16], qui discute également de pistes de réponses, raisonnant notamment sur des ordres partiels plutôt que totaux, ou sur l'introduction d'une incertitude de valuation.

4.2 Alignement de valeurs, environnement éthiques complexes et apprentissage par renforcement multi-objectif

La croissance de l'automatisation des tâches dans des domaines critiques grâce à l'IA rend primordial l'assurance que les agents autonomes agissent en suivant des valeurs proches des humains (value-aligned) [16, 22, 30]. Cette recherche d'agents value-aligned a eu pour conséquence la croissance de l'utilisation d'apprentissage par renforcement dans le domaine de l'IA éthique. En effet, le RL permet d'introduire de l'éthique lors de la modélisation de l'environnement, en introduisant des contraintes éthiques, ou de la modélisation de la fonction de récompense, en y incorporant des variables éthiques (par exemple, l'approche d'ethics-shaping présentée dans [14]). De plus, le RL permet de créer des environnements complexes, proches de la réalité qui peuvent facilement simuler des cadres de prises de décision éthiques.

L'utilisation de plusieurs objectifs lors de l'apprentissage nous permet plus de liberté et de complexité dans la modélisation de la prise de décision d'un agent. D'abord, pour modéliser l'automatisation d'une tâche suivant les valeurs éthiques humaines, plutôt que d'agréger les deux en un seul objectif, il est préferable de considérer d'un côté l'objectif correspondant à la tâche à accomplir, et de l'autre celui correspondant à la dimension éthique de la décision. Ensuite, le choix d'un humain dans une situation donnée, prend en compte plusieurs facteurs éthiques différents et l'utilisation d'un apprentissage multi-objectif nous permet de modéliser chacun indépendamment avec leur propre fonction objectif.

4.3 La question de l'expicabilité de systèmes de décision éthiques

L'intégration de l'éthique dans l'IA, le développement des systèmes de décisions éthiques est motivé par la volonté d'assurance de la morale, de la justice de certaines tâches, amenées à être automatisées. Mais pour s'assurer de la valeur morale, éthique ou juste d'une action ou d'un comportement, il faut comprendre les choix du système. Il faut comprendre précisément pourquoi, dans une situation donnée, une décision est prise plutôt qu'une autre. La question de

l'explicabilité ne pose pas forcément de problème lorsque le modèle sur lequel on raisonne fonctionne avec des contraintes éthiques (rule-based), car les limites éthiques sont définies explicitement. Lorsque l'on s'intéresse à une approche basée sur les données, en revanche, cela peut être un vrai challenge d'apporter une explication des décisions, sans passer par une étude des résultats à posteriori. Tout particulièrement si l'on utilise des réseaux de neurones, dont la prise de décision dépend de plusieurs milliers de paramètres cachés dont la fonctionnalité est implicite. L'explicabilité du Deep Learning est un domaine à part entière et pour le moment, un modèle de prise de décision éthique utilisant des réseaux de neurones ne peut fournir qu'une explication de ses résultats à posteriori. Ce qui ne peut assurer le comportement du système dans le cadre général.

5 Positionnement du sujet par rapport à l'existant

5.1 Articles sur l'intégration de l'éthique dans l'IA

Pour effectuer un état de l'art, j'ai dans un premier temps lu plusieurs survey ([15], [18]) sur l'intégration de l'éthique dans l'IA, présentant les différentes théories représentées et les technologies qui en découlent. Ces survey classes l'ensemble des travaux à ce sujet en deux grandes catégories : les approches rule-based, et data-driven.

Les premières se basent sur une modélisation explicite des règles ou contraintes éthiques lors de la création du modèle et cela peut se traduire par une définition à la main des valeurs éthiques des actions ([8]). Cela peut aussi se traduire par un "contexte éthique" auquel les actions choisies par l'agent doivent répondre ([26]) ou encore une encapsulation totale du MDP : en partant d'un MOMDP prenant en compte les considérations éthiques, calculer un MDP éthique contraint ([25]), de sorte à ce que tous les agents résolvant ce MDP agissent obligatoirement de manière éthique.

Les approches data-driven, quant à elles, se basent sur l'hypothèse que les humains agissent généralement de manière éthique. Collecter un ensemble de choix d'humains va permettre d'apprendre de manière supervisée à des agents à se comporter en suivant les normes éthiques. Cerains modèles modifient directement la fonction de récompense pour y incorporer un terme d'"ethics shaping" prenant en compte les considérations morales de l'agent ([14]). D'autres utilisent des algorithmes multi-armed bandit pour orchestrer les actions de l'agent, lui permettant d'agir selon une politique éthique lorsque c'est necessaire, ou une politique ne prenant pas en compte ces considérations lorsqu'il n'y a pas de risque ([17]). Une approche possible est d'apprendre les contraintes éthiques cachées derrière les comportements humains, pour ensuite apprendre à l'agent à remplir sa tâche en prenant en compte ces contraintes ([27]). L'approche qui a le plus retenu mon attention est celle de [30], qui se base sur les ensembles de trajectoires de plusieurs agents experts des objectifs éthiques, pour ensuite approximer les fonctions objectifs de chacun d'entre eux, et finalement trouver un compromis entre ces fonctions objectifs en approximant les poids d'une combinaison linéaire avec un algorithme de preference-based learning.

5.2 MORL

Plusieurs approches rule-based ([25, 29, 21]) ou data-driven ([30, 17]) ont choisi une approche Multi-Objectif pour modéliser leur environnement (MOMDP), j'ai donc dû me documenter plus précisément sur le sujet, étudier quelles approches étaient les plus adaptées pour le stage. Un des articles que j'ai lu ([28]) est très complet, présentant les différentes architectures de MORL possibles, les avantages et inconvénients de chacune et qui discute de l'évaluation de performance dans le cadre d'algorithmes de résolution de MOMDP.

5.3 Preference-based learning

Lorsque l'on cherche à approximer des poids ou des fonctions inconnus, comme c'est le cas dans les systèmes de décision éthiques, une piste intéressante est l'apprentissage par préférence. Le système pose des questions à un expert, lui demande sa préférence parmi deux actions, dans un état du MDP (ou MOMDP) donné. A partir de sa réponse, on va retirer la partie de l'espace de recherche où les poids (ou la fonction) ne peut plus se trouver, et ainsi jusqu'à avoir une approximation suffisamment précise. Les algorithmes de preference-based learning sont nombreux et ils divergent souvent dans la manière dont on va chercher la question à poser à l'expert (on cherche à réduire le plus rapidement possible l'espace de recherche), et la façon dont on va considérer les réponses de l'expert. On peut considérer qu'il doit toujours avoir une réponse à la question que l'on pose et que sa réponse est absolue, c'est-à-dire qu'il ne peut pas se tromper ou être incertain, mais on peut aussi considérer que l'ordre entre les actions n'est pas total ([4]) et que l'expert n'a pas toujours de réponse. Dans le deuxième cas, on doit raisonner sur des ordres partiels et il n'est alors pas forcément nécessaire d'essayer de départager toutes les actions d'un état, ce qui nous permet une execution plus rapide. Dans le premier cas, la façon dont on va calculer la meilleure question à poser pour faire une dichotomie de l'espace de recherche peut passer par une approximation des hyper-espace des questions potentielles ([30]), ou un calcul d'une valeur d'information de chaque question (AEUS) ([5]). Certains articles récents ([10]) prouvent que ces modèles permettent de résoudre des problèmes classiques de DRL avec des performances proches des algorithmes d'état de l'art du domaine.

5.4 IRL et GANs

Parmi les modèles de MORL dans un cadre éthique, les approches dites "data-driven" se basent sur des données en entrées, censées être des données issues de comportements humains (mais souvent simulées). Ces approches ont donc besoin d'un système pour transformer ces données en des outils utilisables par nos agents RL pour résoudre le MDP (ou MOMDP). Ce que vont faire la plupart des approches ([30, 27, 17]) c'est essayer d'approximer les contraintes cachées ou fonctions de récompenses cachées des humains à partir des choix qu'ils ont fait dans les données. On utilise ensuite ces paramètres approximés pour modéliser le MDP (ou MOMDP) et ensuite apprendre à un agent RL à agir dans cet environnement. Les deux méthodes les plus utilisées pour réaliser cette tâche d'approximation de contraintes/fonctions de récompense cachées sont l'Inversed Reinforcement Learning (IRL) et les Generative Adversarial Networks (GANs).

La première consiste à estimer les paramètres d'une fonction de distribution de probabilité (résponsable de la génération des données expertes) en résolvant un problème de maximum likelihood estimation, souvent en maximisant la log likelihood ([3, 23, 24]).

La seconde méthode ressemble à la première, et peut même parfois être équivalente ([9, 11]), à la différence près que l'on apprend à un autre agent à classifier les données expertes des données générées par le premier agent. Chaque agent apprend en fonction de l'autre, l'agent classifieur cherchant à reconnaître le type de données et l'agent générateur à faire déjouer le premier ([6, 23]).

5.5 Deep L / Deep RL

Bien que le sujet du stage ne porte pas directement sur le deep learning, les méthodes que j'ai recontré en réalisant l'état de l'art, notamment celles des approches data-driven (IRL, GANs et preference-based learning), se basent sur des technologies de deep learning ([4, 5, 10, 24, 30]). On peut également noter que l'article [28], présentant un état de l'art assez complet des modèles et architectures MORL indique que la plupart de ceux-ci utilisent du deep learning. Les approches data-driven étant celles qui ont le plus retenu mon attention, la direction du stage se dirige plutôt vers une utilisation de ces technologies.

5.6 Explainability

Comme présenté dans plusieurs articles ([15, 7]), l'accroissement de l'utilisation de l'IA dans beaucoup de domaines sensibles (justice, transports autonomes, prêts de banques, etc) pousse la recherche dans le domaine à fournir une plus grande transparence et explicabilité des choix des modèles implémentés. Dans un cadre éthique, comme celui du stage, ce sujet a une grande importance, mais la volonté d'explicabilité n'est pas toujours évidente. Particulièrement lorsque beaucoup de modèles MORL du domaine utilisent des réseaux de neurones, technologie dont la complexité rend la difficulté d'explicabilité un de ses principaux désavantages. En effet, les articles portant sur l'explicabilité des réseaux de neurones ([19]), comportent exclusivement sur l'explication des résultats à posteriori. Les techniques employées ne pourront donc jamais garantir un risque nul d'infraction des contraintes éthiques de notre cadre.

5.7 Divers

L'article [16], discute de pourquoi raisonner avec des fonctions de coûts (ou de récompense) n'était pas une bonne approche pour faire de l'apprentissage dans un cadre éthique. Ces fonctions de coûts ou récompense induisant un ordre total sur les actions de chaque état du MDP (ou MOMDP), cet ordre total allant à l'encontre de certains axiomes des théorème de l'impossibilité ou de l'incertitude (un peu à la manière du théorème d'Arrow pour les votes). L'article discute ensuite de comment ne pas aller à l'encontre de ces théorèmes, c'est à dire en raisonnant avec des ordres partiels plutôt que totaux, ou en introduisant de

l'incertitude dans la valuation de nos actions (l'agent n'est pas obligé de savoir s'il considère une action meilleure qu'une autre).

L'article [21] présente un modèle de résolution de dilemmes moraux (dilemme du tramway) à l'aide d'un système de vote inspiré du vote de Nash. Cet article est en lien avec le précédent, car les agents qui vont voter représentent chacun une définition de l'éthique différente (ici déontologie et utilitarisme). Chaque définition ayant un système de valuation des actions potentiellement incomparables, les chercheurs ont décidé de créer ce système de vote qui ne va pas se baser sur une agrégation (combinaison linéaire) des valuations des action par les agents.

6 Travaux effectués lors des deux premiers mois

Les travaux que j'ai réalisés lors de ces premiers mois sont surtout en rapport avec la bilbliographie, la compréhension des différents modèles, les avantages et inconvénients de chacun, lesquels étaient les plus adaptés à la problématique du stage, la compréhension des théories mathématiques sur lesquelles les modèles d'appuient (réseaux de neurones avec descente de gradient, backpropagation, algorithme PPO, IRL et GANs avec les problèmes d'entropie maximum, de cross entropie, de log vraissemblance, etc).

J'ai également pris en main le code de l'article [30], qui a particulièrement retenu mon attention, afin de mieux comprendre leur modèle et les détails de chaque phase du processus. J'ai ensuite créé une nouvelle architecture de code afin de pouvoir lancer le processus complet sur le cluster du lip6, dans l'idée de reproduire les résultats de l'article.

Après avoir étudié en détails le code de l'article, plusieurs questionnements sur l'architecture de leur modèle, font prendre au stage une direction plutôt vers une transformation/adaptation de la phase de preference-learning et une simplification de la phase d'airl. Pour être plus précis, MORAL est une approche data-driven d'intégration de l'éthique dans l'IA, c'est-à-dire que l'hypothèse de base du modèle est que les humains agissent de manière morale et que l'on va se baser sur des données réelles humaines pour faire du RL (MORL en l'occurence). Ces données réelles sont souvent simulées et c'est également le cas ici, elles prennent la forme d'un ensemble de trajectoires d'agents experts éthiques, chacun étant expert en un objectif en particulier, et ils représenteraient les comportements éthiques humains. Cela induit l'hypothèse forte que l'on peut avoir un comportement humain ne se focalisant que sur un objectif éthique, et cela pour tous les objectifs. Cette hypothèse n'est pas du tout triviale car on peut penser que les décisions humaines sont l'agrégation de beaucoup de considérations éthiques différentes et qu'il n'est pas forcément possible de toutes les séparer ou que certaines n'ont pas de sens toutes seules. De plus, collecter des données réelles correspondant à des ensembles de trajectoires complètes d'humains dans un environnement donné, ne paraît pas simple à mettre en place. Ce qui complique l'applicabilité d'un tel modèle dans un cadre réel.

Ces questionnements et réflexions orientent plutôt le stage vers une simplification, ou un retrait complet de cette première phase (AIRL), au profit d'une amélioration de la phase de preference-learning/MORL. On pourrait en effet imaginer un modèle où les données réelles (simulées) seraient introduites au moment de poser les questions à l'expert. Dans un cadre réel, cela se traduirait

par des humains qui remplissent des questionnaires éthiques (on peut penser à un rapprochement avec Moral Machine) où on les place dans plusieurs situations différentes et répondent s'ils préfèrent une action à une autre, ou les résultats d'une trajectoire à ceux d'une autre. Collecter des données, avec un tel questionnaire, paraît réalisable, il faudrait ensuite réfléchir à l'intégration des données dans le système. En effet, dans MORAL, le système essaye de faire une dichotomie de l'espace de recherche des poids possibles pour l'agrégation des fonctions de récompense, en considérant que la réponse de l'expert est "absolue", c'est-à-dire qu'il ne peut pas se tromper ni être incertain. On pourrait améliorer ce système en enrichissant les réponses possibles de l'expert, par exemple une note d'assurance de la réponse qui simulerait à quel point l'expert est certain que la solution a est meilleure que la b. On peut aussi imaginer enrichir les réponses possibles de l'expert, qui pourrait répondre autre chose que sa préférence, par exemple s'il ne faut absolument pas choisir une des deux options ou au contraire si une des deux options est excellente. Ces réflexions de sont que des directions possibles et peuvent être modifiées dans la suite du stage.

7 Calendrier prévisionnel des tâches restant à effectuer

1. Reproduire les résultats de l'article [30].
   dates : 16/05 - 30/05

2. Théoriser un nouveau modèle en s'inspirant de l'article : par exemple changer la partie preference-based learning en trouvant un bon système de méta-heuristique éthique de la combinaison linéaire des fonctions objectifs des experts.
   dates : 30/05 - 13/06

3. Introduire de l'incertitude dans les réponses cet agent expert pour respecter les principes éthiques de [16].
   dates : 13/06 - 20/06

4. Implémenter ces changements de modèle à partir du code existant.
   dates : 20/06 - 18/07

5. Trouver de bons environnements éthiques pour tester notre modèle.
   dates : 18/07 - 25/07

6. Tester notre modèle sur les différents environnements et selon les différentes experts méta-heuristique éthique choisis.
   dates : 25/07 - 15/08

7. Rédiger le rapport final et préparer la soutenance.
   dates : 15/08 - 31/08

References

[1] Thorsten Joachims. "A support vector method for multivariate performance measures". In: *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*. Ed. by Luc De Raedt and Stefan Wrobel. Vol. 119. ACM International Conference Proceeding Series. ACM, 2005, pp. 377–384. DOI: 10.1145/1102351.1102399. URL: https://doi.org/10.1145/1102351.1102399.

[2] Ioannis Tsochantaridis et al. "Large Margin Methods for Structured and Interdependent Output Variables". In: *Journal of Machine Learning Research* 6 (2005), pp. 1453–1484. ISSN: 15337928. URL: http://jmlr.org/papers/v6/tsochantaridis05a.html.

[3] Brian D. Ziebart et al. "Maximum Entropy Inverse Reinforcement Learning." In: *AAAI*. Ed. by Dieter Fox and Carla P. Gomes. AAAI Press, 2008, pp. 1433–1438. ISBN: 978-1-57735-368-3. URL: http://dblp.uni-trier.de/db/conf/aaai/aaai2008.html#ZiebartMBD08.

[4] Weiwei Cheng et al. "Preference-Based Policy Iteration: Leveraging Preference Learning for Reinforcement Learning". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I*. Ed. by Dimitrios Gunopulos et al. Vol. 6911. Lecture Notes in Computer Science. Springer, 2011, pp. 312–327. DOI: 10.1007/978-3-642-23780-5_30. URL: https://doi.org/10.1007/978-3-642-23780-5%5C_30.

[5] Riad Akrour, Marc Schoenauer, and Michèle Sebag. "APRIL: Active Preference Learning-Based Reinforcement Learning". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Vol. 7524. Lecture Notes in Computer Science. Springer, 2012, pp. 116–131. DOI: 10.1007/978-3-642-33486-3_8. URL: https://doi.org/10.1007/978-3-642-33486-3%5C_8.

[6] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[7] Stuart Russell et al. "Letter to the Editor: Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter". In: *AI Mag.* 36.4 (2015), pp. 3–4. DOI: 10.1609/aimag.v36i4.2621. URL: https://doi.org/10.1609/aimag.v36i4.2621.

[8] David Abel, James MacGlashan, and Michael L. Littman. "Reinforcement Learning as a Framework for Ethical Decision Making". In: *AI, Ethics, and Society, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*. Ed. by Blai Bonet et al. Vol. WS-16-02. AAAI Technical Report. AAAI Press, 2016. URL: http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12582.

[9] Chelsea Finn et al. “A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models”. In: *CoRR* abs/1611.03852 (2016). arXiv: 1611.03852. URL: http://arxiv.org/abs/1611.03852.

[10] Paul F. Christiano et al. “Deep Reinforcement Learning from Human Preferences”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 4299–4307. URL: https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

[11] Justin Fu, Katie Luo, and Sergey Levine. “Learning Robust Rewards with Adversarial Inverse Reinforcement Learning”. In: *CoRR* abs/1710.11248 (2017). arXiv: 1710.11248. URL: http://arxiv.org/abs/1710.11248.

[12] Avinash Balakrishnan et al. “Using Contextual Bandits with Behavioral Constraints for Constrained Online Movie Recommendation”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 5802–5804. DOI: 10.24963/ijcai.2018/843. URL: https://doi.org/10.24963/ijcai.2018/843.

[13] Dan Ventura and Darin Gates. “Ethics as Aesthetic: A Computational Creativity Approach to Ethical Behavior”. In: *Proceedings of the Ninth International Conference on Computational Creativity, ICCC 2018, Salamanca, Spain, June 25-29, 2018*. Ed. by François Pachet, Anna Jordanous, and Carlos León. Association for Computational Creativity (ACC), 2018, pp. 185–191. URL: http://computationalcreativity.net/iccc2018/sites/default/files/papers/ICCC%5C_2018%5C_paper%5C_47.pdf.

[14] Yueh-Hua Wu and Shou-De Lin. “A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 1687–1694. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16195.

[15] Han Yu et al. “Building Ethics into Artificial Intelligence”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 5527–5533. DOI: 10.24963/ijcai.2018/779. URL: https://doi.org/10.24963/ijcai.2018/779.

[16] Peter Eckersley. “Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function)”. In: *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*. Ed. by Huáscar Espinoza et al. Vol. 2301. CEUR Workshop Proceedings. CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2301/paper%5C_7.pdf.

[17] Ritesh Noothigattu et al. “Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 6377–6381. DOI: 10.24963/ijcai.2019/891. URL: https://doi.org/10.24963/ijcai.2019/891.

[18] Francesca Rossi and Nicholas Mattei. “Building Ethically Bounded AI”. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 9785–9789. DOI: 10.1609/aaai.v33i01.33019785. URL: https://doi.org/10.1609/aaai.v33i01.33019785.

[19] Richard Meyes, Moritz Schneider, and Tobias Meisen. “How Do You Act? An Empirical Study to Understand Behavior of Deep Reinforcement Learning Agents”. In: *CoRR* abs/2004.03237 (2020). arXiv: 2004.03237. URL: https://arxiv.org/abs/2004.03237.

[20] Alexandra Dufour. *Problèmes éthiques dans l'Apprentissage par Renforcement Rapport de stage-M2 ANDROIDE*. 2021.

[21] Adrien Ecoffet and Joel Lehman. “Reinforcement Learning Under Moral Uncertainty”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2926–2936. URL: http://proceedings.mlr.press/v139/ecoffet21a.html.

[22] Dan Hendrycks et al. “Aligning AI With Shared Human Values”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=dNy%5C_RKzJacY.

[23] Firas Jarboui and Vianney Perchet. “Offline Inverse Reinforcement Learning”. In: *CoRR* abs/2106.05068 (2021). arXiv: 2106.05068. URL: https://arxiv.org/abs/2106.05068.

[24] Shehryar Malik et al. “Inverse Constrained Reinforcement Learning”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 7390–7399. URL: https://proceedings.mlr.press/v139/malik21a.html.

[25] Manel Rodriguez-Soto, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. “Multi-Objective Reinforcement Learning for Designing Ethical Environments”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. Ed. by Zhi-Hua Zhou. ijcai.org, 2021, pp. 545–551. DOI: 10.24963/ijcai.2021/76. URL: https://doi.org/10.24963/ijcai.2021/76.

[26] Justin Svegliato, Samer B. Nashed, and Shlomo Zilberstein. "Ethically Compliant Sequential Decision Making". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 11657–11665. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17386.

[27] Arie Glazier et al. "Learning Behavioral Soft Constraints from Demonstrations". In: *CoRR* abs/2202.10407 (2022). arXiv: 2202.10407. URL: https://arxiv.org/abs/2202.10407.

[28] Conor F. Hayes et al. "A practical guide to multi-objective reinforcement learning and planning". In: *Auton. Agents Multi Agent Syst.* 36.1 (2022), p. 26. DOI: 10.1007/s10458-022-09552-y. URL: https://doi.org/10.1007/s10458-022-09552-y.

[29] Emery A. Neufeld. "Reinforcement Learning Guided by Provable Normative Compliance". In: *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, ICAART 2022, Volume 3, Online Streaming, February 3-5, 2022*. Ed. by Ana Paula Rocha, Luc Steels, and H. Jaap van den Herik. SCITEPRESS, 2022, pp. 444–453. DOI: 10.5220/0010835600003116. URL: https://doi.org/10.5220/0010835600003116.

[30] Markus Peschl et al. "MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning". In: *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*. Ed. by Piotr Faliszewski et al. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022, pp. 1038–1046. URL: https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1038.pdf.