L'inadéquation du " Trolley Problem " pour les véhicules autonomes et l'IA -Working Paper

Marie d'Audibert Caille Du Bourguet

▶ **To cite this version:**

Marie d'Audibert Caille Du Bourguet. L'inadéquation du " Trolley Problem " pour les véhicules autonomes et l'IA -Working Paper. 2025. ⟨hal-05577605⟩

HAL Id: hal-05577605

https://hal.science/hal-05577605v1

Preprint submitted on 2 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

Titre : L'inadéquation du « Trolley Problem » pour les véhicules autonomes et l'IA – Working Paper

Date : 17/01/2025

Auteur : Marie d'Audibdert Caille du Bourguet

Affiliation : Université Cöte d'Azur, CRHI

Résumé :

Cet essai reprend l'argument classique du dilemme du tramway en tant que cadre inadéquat pour comprendre les défis éthiques posés par les véhicules autonomes et les systèmes d'IA contemporains. Même si l'expérience de pensée structure les débats sur le raisonnement utilitaire et déontologique, sa structure binaire ne parvient pas à saisir la complexité, la contextualité et le dynamisme inhérent des environnements du monde réel dans lesquels fonctionnent les systèmes autonomes. Comme l'argument le mentionne, « cette expérience ne pensée ne suffit pas à appréhender les complexités et els réalités des technologies autonomes modernes ». Contrairement à l'agent moral humain imaginé dans le problème du tramway, les véhicules autonomes ne font pas de choix moraux : ils exécutent des instructions programmées, façonnées par des concepteurs qui, eux, sont bien humains. Les systèmes de transports automatisés existants ; tels que les métros et les tramway ; démontrent déjà que la sécurité peut être assurée par la détection d'obstacles et le freinage automatique plutôt que par un calcul moral. Comme le souligne ce point du document, « Les programmeurs peuvent concevoir ces véhicules pour qu'ils s'arrêtent avant de heurter des obstacles, y compris la nécessité de prendre des décisions du type du Trolley Problem ». Cet essais préconise de déplacer le débat éthique des dilemmes hypothétiques vers des enjeux concrets tels que la programmation, la transparence, la réduction des biais, la responsabilité et la souveraineté interdisciplinaire. L'essais conclut que la responsabilité éthique incombe en définitive aux concepteurs, aux opérateurs et aux régulateurs humains, et non aux machines elles-mêmes.

Mots-clés :

Trolley Problem, Ethique, Intelligence Artificielle, Vehicules Autonomes, Responsabilité, Biais, Algorithmes, Gouvernance Technologique, Souveraineté Humaine, Philosophie Morale Appliquée

L'inadéquation du « Trolley Problem » pour les véhicules autonomes et l'IA

Le Trolley Problem est depuis longtemps un élément essentiel des discussions éthiques, en particulier dans le contexte de l'IA et des véhicules autonomes. Le Trolley Problem est né au milieu du XXe siècle, introduit par la philosophe Philippa Foot en 1967. Il a été développé par Judith Jarvis Thomson en 1976. Le problème présente un dilemme moral impliquant un tram en fuite se dirigeant vers cinq personnes sur une voie. Un spectateur doit décider s'il doit tirer un levier, détournant le chariot sur une autre voie où il tuera une personne à la place. Ce scénario soulève des questions sur l'utilitarisme (maximisation du bien général) par rapport à l'éthique déontologique (respect des règles morales quels que soient les résultats). L'avènement de l'IA et des véhicules autonomes a suscité un regain d'intérêt pour le Trolley Problem comme moyen d'explorer la manière dont les machines pourraient gérer les dilemmes moraux. Les véhicules autonomes, en particulier, doivent prendre des décisions rapides dans des environnements complexes, ce qui peut impliquer des scénarios de vie ou de mort. Le Trolley Problem sert d'expérience de pensée pour examiner comment les systèmes d'IA devraient être programmés pour donner la priorité aux vies dans les situations critiques. Le scénario soulève des questions sur les vies qui devraient être valorisées davantage et sur quelles bases, comme l'âge, le nombre de personnes ou les rôles dans la société. Le scénario met en évidence la responsabilité éthique des concepteurs et des programmeurs dans la création d'algorithmes de prise de décision. Il pose la question de savoir s'il est possible ou éthique d'encoder des décisions morales dans des machines et qui devrait être tenu responsable des résultats de ces décisions.

Cependant, cette expérience de pensée ne suffit pas à appréhender les complexités et les réalités des technologies autonomes modernes. Voici pourquoi.

1. **Le Trolley Problem : un bref aperçu** :

Le problème du tramway présente un dilemme moral impliquant un tramway fou qui se dirige vers cinq personnes sur une voie. Une personne doit décider si elle doit tirer un levier, détournant le tramway sur une autre voie où il tuera une personne à la place. Ce scénario implique un humain faisant un choix moral difficile et assumant la responsabilité du résultat. Le problème du tramway est une expérience de pensée en éthique et en philosophie morale, introduite pour la première fois par la philosophe Philippa Foot en 1967. Il est depuis devenu

un élément incontournable des discussions sur les dilemmes moraux, l'éthique et la prise de décision.

La version classique du problème du tramway présente un scénario impliquant un tramway (ou un tramway) qui part en vrille et qui se dirige vers cinq personnes attachées à une voie et incapables de bouger. Un passant, qui se trouve à proximité d'un levier, est confronté à un dilemme moral : tirer sur le levier détournera le tramway sur une autre voie, où il tuera une personne au lieu des cinq.

Le spectateur doit choisir entre deux actions qui entraînent toutes deux des dommages. Le dilemme oblige l'individu à peser les conséquences de sa décision, en se demandant s'il doit agir de manière à provoquer un décès mais sauver cinq vies, ou s'abstenir d'agir, entraînant la mort de cinq personnes. D'un point de vue utilitariste, tirer le levier serait considéré comme le choix moralement juste, car il minimise le dommage global en sauvant le plus grand nombre de vies. Au contraire, un déontologue pourrait soutenir que tirer le levier est moralement mauvais car cela implique de causer activement du tort à une personne innocente, en violation de la règle morale interdisant de tuer.

Le Trolley Problem met l'accent sur l'action humaine et le poids de la responsabilité morale. Le spectateur doit prendre activement une décision et accepter la responsabilité du résultat, qu'il s'agisse d'une action ou d'une inaction. Cette expérience de pensée explore également nos intuitions morales et la façon dont nous priorisons différents principes éthiques lorsque nous sommes confrontés à des choix difficiles. Elle met les individus au défi de réfléchir à leurs valeurs et au raisonnement qui sous-tend leurs jugements moraux. Différentes variantes du Trolley Problem sont utilisées pour examiner un large éventail de cadres éthiques et de théories morales, notamment l'utilitarisme, la déontologie, l'éthique de la vertu, etc.

2. Limitations du Trolley Problem :

Le problème du tramway présente un choix binaire simpliste qui ne reflète pas la complexité des décisions morales de la vie réelle. Il présente un choix binaire brutal entre deux actions distinctes : détourner le tramway pour sauver cinq personnes aux dépens d'une seule personne, ou ne rien faire et laisser le tramway tuer les cinq personnes. Cette nature binaire simplifie à outrance la prise de décision morale, en ignorant les nuances et la complexité des scénarios de la vie réelle. En réalité, les décisions morales impliquent souvent

un éventail d'options et de considérations, plutôt qu'un choix strict du type « l'un ou l'autre ». Par exemple, il peut y avoir d'autres moyens d'atténuer les dommages ou des facteurs supplémentaires à prendre en compte, tels que le contexte, les intentions derrière les actions et le potentiel de conséquences imprévues. Le problème du tramway impose des contraintes artificielles qui ne reflètent pas la nature dynamique des situations du monde réel. Dans la pratique, les individus peuvent rechercher des solutions créatives ou essayer de modifier le scénario pour éviter tout dommage, plutôt que de se limiter aux choix donnés.

Les décisions morales humaines sont souvent beaucoup plus nuancées et dépendent du contexte que le scénario présenté par le Trolley Problem. Les décisions morales humaines sont profondément nuancées et dépendent du contexte . Contrairement au scénario simplifié du Trolley Problem, les dilemmes moraux de la vie réelle impliquent de multiples variables, notamment les relations personnelles, les valeurs culturelles et les conséquences à long terme. Le Trolley Problem fait abstraction des dimensions émotionnelles et psychologiques de la prise de décision. En réalité, les émotions, les expériences et les états psychologiques des individus jouent un rôle important dans la formation de leurs jugements et actions moraux. Le Trolley Problem explore principalement les cadres éthiques utilitaires et déontologiques. Cependant, la prise de décision morale nécessite souvent une synthèse de plusieurs théories éthiques, notamment l'éthique de la vertu, l'éthique du soin et le relativisme moral, pour tenir compte des divers aspects de la vie humaine. Les décisions morales du monde réel nécessitent une sensibilité au contexte et aux facteurs situationnels. Par exemple, la même action peut être jugée différemment selon les circonstances spécifiques, les individus impliqués et les normes sociétales en jeu. Contrairement au scénario statique du Trolley Problem, les situations de la vie réelle sont dynamiques et évolutives. De nouvelles informations et des circonstances changeantes peuvent modifier le paysage moral, obligeant les individus à réévaluer et adapter continuellement leurs décisions.

3. Systèmes automatisés existants :

Les systèmes automatisés comme le métro et le tramway sont déjà opérationnels et programmés pour s'arrêter si un obstacle est détecté. Ces systèmes fonctionnent sur des trajectoires fixes, ce qui facilite la gestion et la prévision des problèmes potentiels. Les systèmes automatisés de métro et de tramway sont en service depuis plusieurs années, ce qui

constitue une étude de cas précieuse pour comprendre les capacités et les limites de la technologie automatisée dans les transports. Ces systèmes sont conçus pour améliorer la sécurité, l'efficacité et la fiabilité des transports urbains.

Les systèmes de métro et de tramway fonctionnent sur des voies fixes, ce qui simplifie leur navigation et réduit la complexité de leur environnement d'exploitation. Ce trajet prédéterminé permet un contrôle et une surveillance précis des véhicules. Ces systèmes suivent un horaire prédéfini avec des points d'arrêt spécifiques, ce qui simplifie encore davantage leur fonctionnement et réduit la variabilité de leurs itinéraires.

Les systèmes de métro et de tramway automatisés sont équipés de divers capteurs, tels que des radars et des caméras, pour détecter les obstacles sur les voies. Ces capteurs surveillent en permanence le chemin à parcourir et peuvent identifier les dangers potentiels, tels que les personnes, les animaux ou les débris. Lorsqu'un obstacle est détecté, le système peut déclencher des mécanismes de freinage automatique pour arrêter le véhicule en toute sécurité, évitant ainsi les accidents et garantissant la sécurité des passagers. Cette capacité est essentielle pour éviter les collisions et maintenir un bon déroulement des opérations. En plus du freinage automatique, ces systèmes sont programmés avec des protocoles d'urgence qui peuvent être activés en réponse à des scénarios spécifiques. Cela inclut l'alerte des opérateurs humains ou des services d'urgence si une intervention immédiate est nécessaire.

Les voies fixes et le fonctionnement programmé des systèmes de métro et de tramway créent un environnement contrôlé, plus facile à gérer et à prévoir par rapport à la nature dynamique et imprévisible du trafic routier. Cette prévisibilité améliore la sécurité et la fiabilité. De nombreux systèmes de métro et de tramway automatisés sont surveillés et contrôlés à partir d'un centre d'exploitation central. Cela permet une surveillance et une intervention en temps réel si nécessaire. Les opérateurs peuvent suivre l'emplacement, la vitesse et l'état de chaque véhicule, garantissant ainsi des performances et une sécurité optimales. Des opérations de maintenance et d'inspection régulières sont effectuées pour garantir que les voies et les véhicules sont en parfait état. Cette approche préventive minimise le risque de défaillances techniques et améliore la sécurité globale du système.

Contrairement aux systèmes de métro et de tramway, les véhicules routiers autonomes évoluent dans un environnement beaucoup plus complexe et variable. Ils doivent naviguer dans des conditions routières diverses, interagir avec d’autres conducteurs, piétons et cyclistes, et s’adapter aux conditions météorologiques changeantes. Les véhicules routiers

autonomes nécessitent des capacités de prise de décision avancées et une flexibilité pour gérer la myriade de scénarios auxquels ils sont confrontés. Cela comprend la planification dynamique d'itinéraire, l'évitement d'obstacles en temps réel et le contrôle adaptatif de la vitesse. Si les principes de détection d'obstacles et de réponse automatisée sont communs, la mise en œuvre pour les véhicules routiers est nettement plus difficile. Les innovations en matière d'IA, d'apprentissage automatique et de technologie des capteurs sont essentielles pour faire progresser les capacités des véhicules routiers autonomes.

4. Véhicules autonomes et complexités du monde réel :

Les véhicules autonomes n'ont pas de pouvoir décisionnel ni la capacité de faire des choix moraux. Ils suivent des instructions programmées créées par des concepteurs humains. Les programmeurs peuvent concevoir ces véhicules pour qu'ils s'arrêtent avant de heurter des obstacles, évitant ainsi la nécessité de prendre des décisions de type Trolley Problem. Les véhicules autonomes évoluent dans des environnements complexes, interagissant avec des conducteurs humains, des conditions météorologiques et des animaux, ce qui diffère considérablement du scénario contrôlé du Trolley Problem.

Contrairement aux humains, les véhicules autonomes n'ont pas la capacité de prendre des décisions morales. Ce ne sont pas des êtres sensibles dotés d'intuition morale ou de capacités de raisonnement éthique. Les véhicules autonomes fonctionnent sur la base d'algorithmes et d'instructions programmés par des concepteurs humains. Ces instructions dictent la manière dont le véhicule doit réagir à divers scénarios, mais elles n'impliquent pas de délibération morale. La responsabilité des actions des véhicules autonomes incombe aux concepteurs et aux ingénieurs humains qui créent et programment ces systèmes. Ils doivent s'assurer que les véhicules fonctionnent de manière sûre et éthique dans le cadre des contraintes de leur programmation.

Les véhicules autonomes sont équipés de capteurs avancés (tels radars et caméras) pour détecter les obstacles sur leur chemin. Ces capteurs analysent en permanence l'environnement et fournissent des données en temps réel au système de contrôle du véhicule. Pour éviter les collisions, les véhicules autonomes peuvent être programmés avec des systèmes de freinage automatique. Lorsqu'un obstacle est détecté, le véhicule peut appliquer automatiquement les freins pour éviter de le heurter. Cette mesure préventive élimine le besoin de décisions morales complexes apparentées au Trolley Problem. Les programmeurs peuvent mettre en

œuvre des mécanismes de sécurité qui donnent la priorité à l'arrêt du véhicule dans des situations incertaines ou potentiellement dangereuses. Ces mécanismes garantissent que le véhicule passe par défaut à un état sûr, réduisant ainsi le risque de blessure.

Les véhicules autonomes évoluent dans des environnements hautement dynamiques, contrairement aux voies contrôlées et prévisibles des systèmes de métro et de tramway. Ils doivent naviguer dans une variété de conditions routières, notamment les embouteillages, les zones de construction et les obstacles inattendus. Les véhicules autonomes partagent la route avec des conducteurs humains, dont le comportement peut être imprévisible. Les véhicules doivent être capables d'anticiper et de réagir aux actions des autres conducteurs, des piétons et des cyclistes. Les conditions météorologiques, telles que la pluie, la neige, le brouillard et la glace, peuvent avoir un impact significatif sur les performances des véhicules autonomes. Les véhicules doivent être capables de s'adapter à ces conditions pour maintenir la sécurité. En plus des autres véhicules, les véhicules autonomes doivent détecter et réagir aux animaux (animaux domestiques et sauvages) et aux piétons. Cela ajoute un autre niveau de complexité à leur fonctionnement, car le comportement des animaux et des piétons peut être erratique. Le fonctionnement réel des véhicules autonomes implique de prendre des décisions complexes basées sur une multitude de facteurs, tels que la vitesse, la distance, les conditions routières et la présence d'autres usagers de la route. Ces décisions sont guidées par la programmation du véhicule et les données qu'il reçoit de ses capteurs.

5. Implications pratiques et éthiques :

L’accent doit être mis sur la programmation éthique et les mesures préventives dans les systèmes autonomes. La programmation éthique implique la conception et le codage de systèmes autonomes de manière à privilégier la sécurité, l’équité et le respect des principes éthiques. Elle garantit que le comportement des véhicules autonomes est conforme aux valeurs sociétales et aux normes morales. L’objectif principal de la programmation éthique est d’assurer la sécurité de tous les usagers de la route. Cela comprend la programmation des véhicules pour qu’ils reconnaissent et réagissent aux dangers potentiels, tels que les piétons, les cyclistes et les autres véhicules. Les protocoles de sécurité doivent être rigoureux et testés de manière approfondie pour minimiser le risque d’accident. Les systèmes autonomes doivent être équipés de mesures préventives pour éviter les situations dangereuses avant qu’elles ne se produisent. Cela comprend des systèmes de freinage automatique , la détection d’obstacles et

des algorithmes de prise de décision en temps réel qui donnent la priorité à l'arrêt ou au ralentissement dans des conditions incertaines. La programmation éthique exige de la transparence dans la manière dont les décisions sont prises et dans les critères utilisés par les systèmes autonomes. Les développeurs doivent être clairs sur les algorithmes et les données qui guident le comportement des véhicules. Cette transparence renforce la confiance du public et garantit que les systèmes autonomes peuvent être tenus responsables de leurs actes. Les programmeurs doivent veiller à éviter les biais dans les données et les algorithmes utilisés pour les systèmes autonomes. Une programmation éthique doit viser l'équité, en veillant à ce que le comportement du véhicule ne discrimine aucun groupe ou individu. Des audits et des mises à jour réguliers peuvent aider à maintenir les normes éthiques et à remédier à tout biais émergent. L'intégration des idées des éthiciens et des experts en philosophie morale peut guider l'élaboration de normes de programmation éthiques. Cette approche interdisciplinaire garantit que les systèmes autonomes reflètent un large éventail de considérations éthiques.

La responsabilité ultime incombe aux concepteurs et aux opérateurs humains de ces systèmes. La responsabilité ultime du comportement des véhicules autonomes incombe aux concepteurs et aux ingénieurs humains qui créent et programment ces systèmes. Ils doivent s'assurer que les principes éthiques sont intégrés dans la conception et que les véhicules fonctionnent de manière sûre et fiable. La supervision humaine est essentielle pour la surveillance et l'amélioration continues des systèmes autonomes. Même après le déploiement, les opérateurs humains doivent superviser les véhicules, réagir aux anomalies et mettre à jour le système si nécessaire. Les concepteurs et les opérateurs doivent se conformer aux normes réglementaires et aux directives établies par les gouvernements et les organismes industriels. Ces réglementations sont en place pour protéger la sécurité publique et garantir le fonctionnement éthique des véhicules autonomes. Il est essentiel d'établir des mécanismes de responsabilisation clairs. En cas d'incident impliquant un véhicule autonome, il devrait être possible de retracer les décisions et les actions prises par le système et d'identifier toute erreur ou omission humaine. Les mécanismes de responsabilisation aident à maintenir la confiance du public et fournissent une base pour les évaluations juridiques et éthiques. Les ingénieurs et les développeurs travaillant sur des systèmes autonomes doivent recevoir une formation à la prise de décision éthique et à l'utilisation responsable de l'IA. Cette formation les aide à comprendre les implications plus larges de leur travail et favorise une culture de sensibilisation éthique au sein de l'industrie. En impliquant les parties prenantes, notamment le public, les décideurs politiques et les groupes de défense, dans le développement et le

déploiement des véhicules autonomes, on s'assure que des points de vue divers sont pris en compte. Cet engagement permet de répondre aux préoccupations de la société et d'aligner la technologie sur les valeurs et les attentes du public.

Le Trolley Problem, bien que très répandu dans les débats éthiques, ne permet pas de répondre aux défis et aux décisions liés à l'IA et aux véhicules autonomes. Ces technologies fonctionnent dans des environnements bien plus complexes et nécessitent une programmation éthique et une responsabilité humaine pour garantir un fonctionnement sûr et éthique.