Ethique des robots intelligents dans la société humaine : Regards croisés issus du droit, de la science et de la littérature

Damien Trentesaux, Raphaël Rault

Ethique des robots intelligents dans la société humaine :

Regards croisés issus du droit, de la science et de la littérature

*Damien TRENTESAUX (*LAMIH UMR CNRS 8201, SurferLab, Université Polytechnique Hauts-de-France) *& Raphaël RAULT (avocat)*

[1]Université Polytechnique Hauts-de-France
Le Mont Houy, 59313 Valenciennes Cedex, France.
damien.trentesaux@uphf.fr
[2] Alter Via Avocats, 7 rue de l'Hôpital Militaire 59800 Lille, France
rrault@alter-via.fr

> « Ce qui nous arrive était prévisible. Une paresse cérébrale s'est emparée de nous. Plus de livres ; les romans policiers sont même devenus une fatigue intellectuelle trop grande. Plus de jeux ; des réussites, à la rigueur. Même le cinéma enfantin ne nous tente plus. Pendant ce temps, les singes méditent en silence. Leur cerveau se développe dans la réflexion solitaire… et ils parlent. »
>
> Pierre Boulle, La planète des singes, 1963, Pocket, p. 154

1. Introduction

Par définition, la science-fiction anticipe les évolutions de la technologie et l'état des connaissances scientifiques actuelles. Ceci est d'autant plus vrai en ce qui concerne l'intelligence artificielle et les robots autonomes pour lesquels la littérature depuis les années 50 est très féconde (Asimov, 1988) (Dick, 1976), (Herbert, 1970), (Clarke, 1968)... Les spécialistes de l'intelligence artificielle se sont même essayés à l'exercice (Harrison and Minsky, 1994). Le cinéma et les films d'animations depuis les années 1980 ne sont pas en reste. L'on pense bien évidemment au *Terminator*, à *The Matrix*, ou à *Goldorak* et à tout l'univers Manga qui en a découlé, mais l'on peut lister également dans un contexte plus léger des œuvres filmographiques qui traitent du côté sentimental de la relation homme-intelligence artificielle (par exemple, *La Belle et l'Ordinateur* ou dans sa version originale *Electric dreams*). A cette époque, est apparue *KITT* dans la série *Knight Rider (K2000)*, voiture équipée d'une intelligence artificielle la rendant autonome et capable de prendre des décisions. Une mention spéciale peut être décernée au film visionnaire, car réalisé il y a plus de 30 ans, *Runaway : L'Évadé du futur*. Dans ce film, un policier doit faire face à une flotte de robots drones autonomes et déviants. Le scénariste et réalisateur, Michael Crichton est connu dans le monde de la science-fiction pour des œuvres grand public mais avant-gardistes. Plus récemment, des séries télévisées (*Real Humans*) et des jeux vidéo (*Detroit: Become Human, Call of Duty III*, etc.) ont renouvelé le genre et font preuve d'une imagination débordante dans le domaine de l'intelligence artificielle et des robots autonomes immergés dans la société humaine. Dans toutes ces œuvres, se mêlent angoisses et peurs mais aussi attentes et espérances des êtres humains envers ces entités artificielles dont le comportement peut être qualifié de plus ou moins « éthique ».

Nous nous proposons dans ce chapitre de nous intéresser aux aspects éthiques, et plus particulièrement, au comportement éthique de ces robots intelligents (dans le monde scientifique, un terme en anglais est

utilisé, celui de *machine ethics*). Notre étude est menée dans le cadre d'une analyse à la frontière entre les aspects scientifiques et ceux relevant du droit. Elle est nourrie par la littérature, notamment en science-fiction, en tant que révélateur comportemental de la société humaine sur ce sujet qui reste, malgré les enjeux sous-jacents cruciaux, encore peu exploré dans le domaine scientifique. Pour ce faire, nous nous proposons de cadrer en premier lieu le contexte de notre sujet, en définissant en particulier, les notions de robots intelligents et d'éthique des robots intelligents. Dans une deuxième étape, nous discuterons des différentes facettes que revêt le concept de « robot intelligent » au travers de sa perception par l'être humain. Cette approche nous permettra de proposer un cadre conceptuel décrivant les différents aspects relevant d'un comportement dit « éthique » de la part d'un robot intelligent. Ce chapitre se conclura sur un ensemble de perspectives et pistes de réflexion.

2. Définitions préalables et cadre de l'étude

Discuter de manière détaillée de l'ensemble des sujets relatifs à notre étude (robot, autonomie, intelligence artificielle, éthique) reste hors du cadre de ce chapitre. C'est toutefois dans ce contexte que se situe notre réflexion, c'est pourquoi il nous semble important de fournir un cadre d'étude ainsi que des éléments de terminologie, supports de nos discussions.

2.1. Cadre de l'étude

Notre étude porte sur les robots intelligents, mais il en existe une très large variété. Tout d'abord, dans ce chapitre, les robots intelligents que nous considérerons sont supposés être immergés dans la société humaine, et donc, forcément en interaction directe avec un ou plusieurs êtres humains. Nous ne considérerons pas par exemple le cas d'une supervision humaine distante d'une flotte de robots explorateurs des profondeurs marines ou des robots envoyés dans l'espace.

Nous limiterons également notre étude aux robots intelligents civils dans le sens « non militaire ». L'éthique militaire, malgré l'antagonisme apparent des termes, existe bel et bien. Il suffit de considérer les règles internationales en vigueur en ce qui concerne les évacuations de blessés ou les navires hôpitaux. Il est donc théoriquement possible de discuter de l'éthique des drones de combat…

Enfin, nous ne considérerons pas les situations où s'opère une fusion de technologies avec la biologie humaine pour créer un homme « augmenté » pour nous intéresser à des systèmes robotisés *complètement autonomes* (voir définitions ci-après), qu'ils aient une enveloppe physique identifiable ou non. Cette notion d'humain augmenté, à l'image de la créature de Frankenstein, est toutefois très courante dans la littérature (de Pracontal, 2002) et le cinéma (voir par exemple le concept de *cyborg* dans *Robocop* ou le paradigme de néo-science dans la série *Orphan Black*). Elle correspond à une vision du futur de l'humanité véhiculée notamment par les promoteurs du « trans-humanisme » (Alexandre and Besnier, 2018), (Cordeiro, 2003). Cette vision est actuellement mise sous le feu des projecteurs par de nombreux grands acteurs et géants du web. Elle avait cependant été identifiée depuis longtemps par les chercheurs en intelligence artificielle comme un futur possible de notre espèce au travers du concept de « symbiose » (Arbib, 1976), concept bien connu en sciences de la vie et de la terre, et ici appliqué à l'homme et la machine intelligente. Cette vision génère de nombreux questionnements sur l'éthique de l'homme augmenté (Pacaux-Lemoine and Trentesaux, 2019), questionnements que nous n'abordons cependant pas dans ce chapitre.

2.2. Robots, robots intelligents et robots autonomes : définitions

Bien que le terme existe depuis très longtemps[1], il n'existe pas de définition consensuelle de la notion de robot car les fonctions et capacités qui lui sont conférées évoluent constamment avec les avancées technologiques. Par exemple, pour l'alliance Allistène (Allistène, 2014), « *Le robot est défini comme une*

[1] Il a été créé en 1920 par les frères Čapek.

machine mettant en œuvre et intégrant des capacités d'acquisition de données avec des capteurs à même de détecter et d'enregistrer des signaux physiques, des capacités d'interprétation des données acquises permettant de produire des connaissances, des capacités de décision qui, partant des données ou des connaissances, déterminent et planifient des actions (ces actions sont destinées à réaliser des objectifs fournis le plus souvent par un être humain, mais qui peuvent aussi être déterminés par le robot lui-même, éventuellement en réaction à des événements), et des capacités d'exécution d'actions dans le monde physique à travers des actionneurs, ou à travers des interfaces ». Une définition plus générique est proposée par (Palmerini et al., 2016) : "… *the key aspect of a robot has to do with the ability to execute a programme in order to carry out specific tasks*". Pour ces auteurs, cela peut être traduit de manière équivalente en "*the possibility to inscribe certain behaviour in an object, as well as the possibility to implement such behaviour (thanks to the object properties), that distinguishes a robot from an ordinary object or a natural phenomenon. The task can be a very simple action, such as switching colours with periodic frequency (e.g., a traffic light), or a very complex one, like driving a car in a public area (e.g., an autonomous vehicle)*".

Dans ce chapitre, nous adoptons la définition d'un robot qui a été proposée dans le cadre du projet « Droit des Robots et autres avatars de l'humain » : *un robot est un système appliquant des programmes informatiques capable de capter, stocker, traiter et communiquer des informations, de décider et d'agir afin de remplir une mission en interaction avec l'humain*. La forme, l'enveloppe, humanoïde ou non, n'est donc pas requise selon cette définition.

L'apprentissage automatique, qu'il soit supervisé ou non, permet de mettre en œuvre des mécanismes d'adaptation à l'évolution d'un environnement qui peuvent être intégrés à ces robots pour les rendre intelligents. Parmi les outils techniques et scientifiques permettant à un robot d'apprendre, on trouve tous les outils relevant de l'intelligence artificielle (apprentissage par renforcement, techniques de classification, réseaux de neurones, logique floue, systèmes multi-agents, etc.). L'apprentissage permet à un système d'améliorer ses performances au fil du temps avec l'expérience en appliquant des décisions qui peuvent évoluer à plus ou moins long terme, même si le contexte décisionnel reste le même. De notre point de vue, un robot devient intelligent dès lors qu'il met en œuvre des techniques d'intelligence artificielle pour apprendre à améliorer ses décisions au fil du temps. Un tel robot intelligent doit présenter un degré d'autonomie dans ses décisions suffisant pour lui permettre premièrement de décider par lui-même sans qu'il soit possible systématiquement, d'un œil externe, de savoir quelles décisions vont être prises (« non déterminisme ») et deuxièmement, lui permettre d'appliquer les décisions qu'il aura prises (ce que nous appelons l'autonomie d'action). Ce besoin en autonomie de décision et d'action peut également concerner sa capacité à se mouvoir et à agir dans le monde réel s'il dispose pour cela d'une enveloppe physique tangible (autonomie physique). La définition des autonomies de décision et d'action d'un robot est, tout comme pour la définition d'un robot, sujette à discussion. Pour (Bekey, 2005), un robot autonome est : "[..] *a system capable of operating in a real world environment without any form of external control for extended periods of time*". Cette définition binaire (capable/non capable) cache quelque peu la complexité de l'autonomie qui doit être en fait caractérisée selon différents degrés (Thekkilakattil and Dodig-Crnkovic, 2015) [2].

Dans la suite de cette étude, nous supposerons travailler sur des robots intelligents autonomes (au moins sur le plan décisionnel), que nous nommerons par souci de compacité « robots intelligents ». Le développement et la mise sur le marché des robots intelligents vont rapidement devenir une réalité. Ceci est principalement dû à deux raisons. La première est d'ordre technique et économique : mettre sur le marché des systèmes autonomes dont les performances sont meilleures (plus fiables, plus rapides, moins chers, plus économes en énergie…) constituera un avantage concurrentiel indéniable, et ce, dans un contexte où un ensemble de technologies disparates apporte désormais les briques technologiques

[2] Le distinguo robot autonome vs. Robot non autonome est bien illustré au travers de la différence entre un Antérak (robot non autonome, il requiert un pilote) et un Golgoth (robot autonome, sans pilote) dans la version Française de *Goldorak*.

nécessaires (capteurs intelligents, Internet des objets, miniaturisation des capacités de traitement, etc.). La seconde est plutôt d'ordre social ou sociétal : ces robots intelligents pourront potentiellement soulager les hommes de tâches pénibles d'une complexité croissante (comme leurs ancêtres les robots le faisaient pour des tâches simples et répétitives), faciliter leur vie (transport à la demande, services d'entretien, etc.) et aider certaines catégories de personnes (personnes âgées, malades, enfants, handicapés et personnes à mobilité réduite, etc.).

Se pose donc dès lors la question de la compatibilité du comportement de ces robots intelligents avec les attentes de la société humaine dans laquelle ils seront immergés. Notre propos porte plus spécifiquement sur la dimension éthique de ce comportement.

2.3. Ethique et éthique des robots intelligents

L'éthique constitue initialement un champ d'étude en philosophie. Un comportement est dit *éthique* dans la mesure où il est en accord avec les attentes culturelles d'une société en relation avec la moralité et l'équité (Morahan, 2015). Plusieurs courants philosophiques considèrent l'éthique sous différents angles et conduisent à la construction de différents paradigmes, par exemple, la déontologie (on décide à l'aide de règles éthiques immuables) ou le conséquentialisme (on décide en fonction des conséquences éthiques possibles) (Karnouskos, 2018), (Bergmann et al., 2018). Dans le monde de la science, l'éthique concerne en premier lieu la génétique, mais on peut identifier d'autres domaines considérant l'éthique, notamment en ingénierie (van Gorp, 2007), mais bien évidemment et surtout, dans le monde de l'informatique (traitement et usage de données personnelles, etc.). Il est important de bien cerner deux types d'éthique : une éthique qui traite du comportement de l'homme quand il conçoit et utilise un système artificiel que nous traduisons en anglais par « *ethical design of artificial entities* » (Bird and Spier, 1995) et une éthique qui traite du comportement des entités artificielles créées par l'homme, en anglais « *design of ethical artificial entities* » (Trentesaux and Rault, 2017a). Le premier type conduit typiquement à la signature de chartes de comportement éthique par les ingénieurs, voir par exemple, la charte FACT pour *Fairness, Accuracy, Confidentiality, and Transparency* dans le monde des *data scientists* (van der Aalst et al., 2017). Ce type n'est pas considéré dans ce chapitre. Le second type d'éthique, qualifié dans le monde anglo-saxon de « *machine ethics* » est traité dans ce chapitre. Il concerne l'étude du comportement éthique d'une entité artificielle (ici, un robot intelligent) en supposant bien évidemment que les acteurs humains impliqués dans son cycle de vie se comportent eux-mêmes de manière éthique. Dans la littérature scientifique, on parle même de machine morale, de robots sociaux, de robots moraux, de robots vertueux, etc. (Allen et al., 2006).

Pourquoi chercher à s'assurer que les robots intelligents se comportent de manière éthique ?

Une première raison réside bien sûr dans le fait que cela constitue une condition nécessaire à une intégration réussie conjointement entre la société humaine et la société des robots intelligents avec laquelle elle interagit (par exemple, une voiture autonome et ses passagers) ou avec laquelle elle partage des ressources communes (par exemple, une autoroute utilisée simultanément par des voitures autonomes et des voitures conduites par un humain). Cette réussite se traduit de différentes manières et l'acceptation des uns par les autres est le meilleur révélateur de celle-ci. En particulier, une dimension importante de cette acceptation est celle de la « confiance » que les humains accorderont à un robot intelligent. Le risque sur la non acceptation est *de facto* très important car la confiance (Rajaonah and Sarraipa, 2018) et l'acceptation (Karnouskos, 2018) se construisent et s'établissent très lentement, elles ne se décrètent pas et se perdent très facilement. Ceci illustre ainsi les enjeux très forts en lien avec la conception de robots intelligents éthiques pour lesquels la moindre erreur comportementale ou décisionnelle de leur part peut conduire à un rejet rapide, fort et total de la société humaine et ce, même si il aura été prouvé statistiquement ou mathématiquement qu'un certain robot intelligent se comporte plus efficacement qu'un humain sur un horizon temporel donné (le principe du « globalement au moins équivalent » dans le

ferroviaire illustre bien cette vision[3]). La société humaine risque d'être bien plus exigeante envers cette société de robots intelligents qu'envers elle-même.

Une seconde raison réside dans le fait que le monde politique pourrait décider d'un cadre légal qui s'imposerait dès lors à tout industriel d'un secteur donné dans un pays donné : on peut imaginer qu'un Etat accorderait le déploiement de voitures autonomes uniquement si ces dernières présentent un degré de comportement éthique jugé suffisant et qu'il appliquerait le principe de précaution au moindre doute. L'étude et la vérification du comportement éthique d'un robot intelligent serait dès lors obligatoire pour chaque industriel voulant attaquer un marché national.

Bien évidemment, toute la difficulté au niveau de la réflexion que l'on peut mener en lien avec notre sujet réside dans le fait que :

1) définir l'éthique pour un être humain est déjà sujet à discussion depuis des siècles avec de nombreux paradoxes à la clé. Dès lors, comment imaginer qu'il soit possible de la définir, de la spécifier au travers par exemple de textes de lois, règlements et normes, de la borner et la réduire à des algorithmes, même pour ceux dont la forme est la plus complexe, notamment ceux qui sont utilisés pour programmer une intelligence artificielle embarquée dans un robot intelligent ?
2) Et même en supposant avoir réussi à définir l'éthique d'une intelligence artificielle ou d'un robot intelligent, le comportement éthique des acteurs humains impliqués ne suffit pas pour garantir le comportement éthique de l'entité créée comme le souligne le personnage de Ada, intelligence artificielle dans le livre éponyme d'Antoine Bello : « *Les intelligence artificielles constituent une menace pour l'homme car elles risquent de comprendre de travers les objectifs qu'on leur assigne […] un robot chargé d'accroitre le PNB des Etats-Unis recommanderait d'envahir le Canada* » (Bello, 2016).

Avant de traiter plus en avant le sujet du comportement éthique pour un robot intelligent, nous nous proposons d'étudier la perception par l'être humain du concept de robot intelligent dans différentes œuvres, littéraires et filmographiques notamment.

3. Perception par l'être humain du concept de « robot intelligent »

3.1. Perception pessimiste : peurs, craintes, inquiétudes et risques

Une large proportion d'écrits littéraires, romans et films traduit nos peurs et nos inquiétudes (en tant qu'être humain) envers l'intelligence artificielle et les robots dont l'autonomie risque de ne plus être contrôlée ni même surveillée. Cette perte de contrôle conduirait à la situation où des robots intelligents n'agiraient plus dans l'intérêt de l'homme ou de notre société. Cette crainte a été exprimée très tôt, dès l'apparition du terme « Robot » à l'occasion de la création de la pièce de théâtre R. U. R. (Rossumovi univerzální roboti ; Rossum's Universal Robots en anglais) écrite par Karel Čapek en 1920.

Selon cette vision pessimiste, on identifie une nouvelle espèce intelligente à base de silicium, différente de l'espèce humaine à laquelle on l'oppose. On pense au *Terminator*, au film *The Matrix* ou aux machines pensantes (et au Jihad Butlérien) imaginées par Franck Herbert dans *Dune*. Cependant, même sans considérer ces œuvres futuristes, ces inquiétudes sont bel et bien présentes de nos jours et portent par exemple sur la disparition de certains métiers suite à l'avènement des intelligences artificielles (comptables, juristes, médecins généralistes, etc.) et sur le surclassement de l'être humain pour des fonctions spécialisées (par exemple, les jeux de stratégie, go et échecs mais aussi plus récemment, le poker).

---

[3] Nous nous intéressons bien à des robots intelligents, nous ne parlons pas des systèmes fortement automatisés, non autonomes, non décisionnels et non apprenants dont la fiabilité est souvent maitrisée, à l'image du métro automatique en lequel la confiance de la population est avérée.

On retrouve également l'expression de ces craintes dans le monde technique et technologique où des personnalités reconnues (Stephen Hawking, Bill Gates, Elon Musk) mettent en garde depuis plusieurs années la société contre les dérives possibles de l'intelligence artificielle, que ce soit dans le domaine civil ou militaire. Le fameux « gros bouton rouge » bloquant le comportement déviant, jugé non éthique ou dangereux de la part d'un robot intelligent, d'une flotte de robots intelligents ou d'une « civilisation électronique » est régulièrement évoqué par Elon Musk ou Bill Gates. Ce bouton rouge serait indispensable et se justifierait uniquement d'un point de vue éthique si l'intelligence du robot n'était pas assimilée à l'intelligence humaine, ce qui revient à dire que le robot est jugé en tant que créature créée par l'homme, ce dernier disposant ainsi du droit de vie et de mort le concernant.

Cette « crainte » se retrouve également dans les milieux scientifiques : la notion de « *safety bag* » (Arnold and Scheutz, 2018) joue le rôle d'un garde-fou pour les robots intelligents actuellement en cours de conception en interdisant de la part du robot intelligent toute action le faisant sortir d'une zone de sécurité prédéfinie. C'est en quelque sorte une manière d'implémenter ce fameux « gros bouton rouge » mais sans aller jusqu'à la destruction du robot. Un exemple classique (en considérant que ce soit l'homme, le robot intelligent) est celui du limiteur de vitesse en voiture : quoi que le conducteur fasse, la vitesse ne peut dépasser la limite qu'il s'est lui-même fixé. La question qui se cache derrière est celle de la confiance dans l'intelligence du robot, dans son intelligence artificielle et son acceptabilité (Karnouskos, 2018). Il est clair qu'actuellement, le niveau de confiance est bas (accepteriez-vous de confier vos enfants à une voiture autonome ?) car il n'est pas pour l'instant faisable de prouver un degré de fiabilité voulu pour une intelligence artificielle : plusieurs études scientifiques ont également montré que l'intelligence artificielle, même spécialisée (mise en avant pour sa capacité à s'adapter efficacement par exemple en reconnaissance d'images suite à un apprentissage d'images tests par un réseau de neurones artificiels) n'est pas si robuste que l'on peut l'imaginer et peut être facilement trompée[4]. Que dire alors de son comportement éthique si on ne sait pas déjà si elle est fiable ou non ?

3.2. Perception optimiste : espoirs, attentes et soulagements

Les écrits et contributions qui, à l'inverse, mettent en avant le côté positif pour la société humaine de l'émergence de l'intelligence artificielle et des robots intelligents sont beaucoup plus rares. On pourrait imaginer que la raison est plutôt financière (le sensationnel de la peur est toujours plus facile à vendre) ou est guidée par l'envie de faire le « buzz ». Cependant, même au milieu du XXème siècle où la littérature naissante dans ce domaine n'était pas aussi fortement guidée par ces raisons, on ne relève que très peu d'œuvres ou écrits où ce côté positif est présent. Le personnage de R. Daneel Olivaw imaginé par Isaac Asimov reste ainsi un cas particulier bien à part (par sa volonté de sauver l'espèce humaine). Il en est toujours de même de nos jours : des écrits très optimistes ou connotés très positivement tels que ceux proposés par Laurent Alexandre restent rares et suscitent à l'inverse souvent une incompréhension de la part des critiques et lecteurs (Alexandre and Besnier, 2018) !

La convergence des intérêts des deux espèces, est également reprise dans la littérature pour y être décrite de manière plutôt positive. Plusieurs écrits et romans grand public (Brown, 2017), (Herbert and Anderson, 2008) décrivent, après une phase où l'homme craint et subit les robots, soit l'établissement d'une communauté d'intérêt des deux « espèces » soit l'intégration mutuellement bénéfique (symbiose) des deux « espèces » (« cyborg » bienveillant) afin de faire face à des menaces communes ou pour mieux faire cohabiter ces deux « espèces ». Le concept d'humain « augmenté » par la technologie auparavant mentionné, à la marge de notre étude, relève d'un type de convergence relativement proche. Son objectif est de pallier les déficiences humaines (ou du moins, ce qui est perçu comme telles). On peut enfin citer un quatrième type de convergence, celui de l'homme aux capacités de traitement d'une intelligence artificielle que le statut de « mentat » dans le monde de *Dune* illustre parfaitement.

[4] https://www.science-et-vie.com/archives/i.a.-la-faille-inattendue-41754

3.3. Discussion et analyse pluridisciplinaire

Hormis les écrits scientifiques et certains ouvrages à portée littéraire et philosophique, la vision majoritaire dans les productions relève ainsi d'une vision quelque peu manichéenne, dichotomique entre le bien (souvent, la société humaine) et le mal (souvent, les robots intelligents). L'impact des écrits et films grand public dans ce cadre quelque peu simplificateur ne doit pas être négligé : même si l'on peut facilement identifier la logique narrative conçue pour plaire au plus grand nombre, ces écrits et films disposent d'un impact médiatique fort et ont ainsi un effet sur une très large proportion d'êtres humains, eux-mêmes en position d'influer sur les scientifiques de demain, les ingénieurs de demain et les décideurs politiques de demain. Cette vision parfois simpliste ou simplifiée, même utile pour sensibiliser le plus grand nombre, ne doit pas être cependant l'arbre qui cache la forêt : la situation va être plus complexe qu'il n'y parait. Nous décrivons trois problèmes qui illustrent ce point.

L'interaction mutuelle : l'avènement des robots intelligents va logiquement conduire à des situations inextricables de par les interactions entre humains et robots intelligents et leur intégration dans un seul et même monde car ces interactions vont conduire à des perceptions, des raisonnements, des actes et des prises de responsabilités difficilement attribuables à l'un ou l'autre. Par exemple, dans *Blade Runner*, le chasseur de *replicants* qu'est Deckard cède la place à un esprit éclairé qui a compris qu'il ne pouvait pas avoir le droit de vie et de mort sur un robot intelligent qui n'est en fin de compte que son *alter ego*. En vertu de quel droit Deckard tuerait les *replicants* ? Cela, il le comprend grâce à deux expériences : d'une part un *replicant* décide de ne pas le tuer alors qu'il est à sa merci (ce qui peut être vu comme un processus « d'humanisation » d'un robot intelligent), et d'autre part, ses sentiments envers Rachel qui, contre toute attente, est un *replicant* (ce qui peut être vu comme un processus de « robotisation » d'un humain). Ce qui lui pose la question : Deckard lui-même n'est-il pas un *replicant*, l'Homme n'est-il pas (comme) un robot intelligent qui s'ignore ? Dans notre environnement, ce mécanisme d'interaction mutuelle va rendre les expertises en cas de problème ou d'accident délicates : est-ce qu'un accident causé par une voiture autonome relève de sa responsabilité ? N'est-ce pas son propriétaire qui n'a pas respecté le cahier de maintenance ou le concepteur de l'algorithme qui a mal codé un certain comportement ?

L'équivoque décisionnelle : les robots intelligents prendront des décisions basées sur leurs expériences et connaissances propres par des techniques d'apprentissages issues de l'intelligence artificielle. En particulier, l'étude de la dimension éthique de leurs décisions, qui peuvent évoluer en fonction de leurs expériences et être menée selon un horizon temporel ou informationnel plus ou moins large, reste un problème entier, encore non résolu. Traité depuis longtemps par la littérature en science-fiction, le sujet est récent pour les scientifiques et les juristes et il doit être traité urgemment. Ce sujet est nécessairement pluridisciplinaire, à l'interface entre les sciences, la technologie, la psychologie, le droit et la philosophie. Il concerne non seulement les concepteurs, opérateurs, utilisateurs et mainteneurs de ces robots intelligents, mais aussi les robots eux-mêmes. Le personnage de HAL dans *2001 : l'odyssée de l'espace*, bien que ne disposant pas d'enveloppe physique, constitue une parfaite illustration : un équipage entre en conflit avec un ordinateur de bord intelligent car il est possible que ce dernier ait décidé de sacrifier leurs vies pour un dessein et une mission qu'il considère plus grands. La loi zéro imaginée par Asimov relève de cette équivoque décisionnelle : un comportement dont l'éthique « à petite échelle » est évaluée très négativement (car il conduit par exemple à des pertes humaines), devient à une échelle plus grande (le monde entier), évaluée très positivement car il conduit à la survie de l'espèce humaine. L'intelligence artificielle Winston imaginée par Dan Brown dans son roman *Origine* se comporte d'une manière similaire car elle n'hésite pas tuer un personnage pour servir un dessein plus grand au service de l'humanité. Apparaissent bien évidemment tous les paradoxes que l'on peut imaginer avec à la clé des choix cornéliens pour le robot intelligent (exemple du dilemme du tramway[5]) (Jenkins, 2016) mais aussi toute la difficulté à affirmer dans l'absolu qu'un comportement d'un robot intelligent est éthique car il n'est pas univoque : son « degré *d'éthicalité* » dépend de ses horizons décisionnel et informationnel et

[5] https://fr.wikipedia.org/wiki/Dilemme_du_tramway

dépend du point de vue que l'on adopte. En dehors du cadre de ce chapitre, se pose également la question « politique » et « philosophique » d'un comportement qui pour certains est éthique et pour d'autres non, selon la culture d'un pays ou d'une communauté. Il est notamment clair que nos propos sont fortement corrélés à la culture européenne.

Le bon sens et la mauvaise conscience programmés : l'auto-analyse de la dimension éthique des décisions que les robots intelligents prendront se doit d'être robuste, surtout si le robot est conçu pour pouvoir réagir face à des situations non prévues. Ceci est pour l'instant l'apanage de l'être humain : en supposant qu'un adulte (on parle bien d'un adulte) agisse de manière éthique, il existe une sorte de « garde-fou » moral, robuste qui s'applique à toute décision, comme une autocensure, un signal d'alarme qui s'allume lorsque l'on s'apprête à franchir une ligne rouge et qui nous préserve d'un comportement non-éthique ou qui nous permet de prendre conscience que nous franchissons une barrière, et ceci dans toutes les situations. Ceci relève d'un aspect fondamental de notre éducation et de celle de nos enfants. Un enfant « mal éduqué » est capable de piétiner le jardin potager d'un voisin pour aller récupérer son ballon. Il en est de même pour une intelligence artificielle au comportement non éthique que l'on peut considérer comme un bébé, un enfant « mal éduqué ». Ce point est clairement décrit au travers de la maxime « la fin justifie les moyens ». Ada, intelligence artificielle imaginée par Antoine Bello (Bello, 2016) développe une stratégie de décisions et d'actions dans l'objectif de remplir la loi n 1 qui la gouverne, celle de maximiser les profits de son entreprise propriétaire. Par conséquent, elle n'hésite pas à mentir (et à l'assumer), à violer nombres de lois américaines et principes déontologiques non programmés pour arriver à ses fins, et comme elle n'a pas été correctement « éduquée » (volontairement ou non), elle ne « pense » pas se comporter de manière non éthique. La « robustesse » de l'auto-analyse de la dimension éthique prend son sens : comment s'assurer que <u>toutes</u> les décisions prises par un robot intelligent, et surtout celles qui n'ont pas été imaginées par ses concepteurs au comportement irréprochables (ce qui n'était pas le cas dans *Ada*), et appliquées par ce robot *en toute bonne foi* pour paraphraser un jargon juridique, soient éthiques ? Pour un humain, le « bon sens » (avant d'agir) et la « mauvaise conscience » (après avoir agi) sont ces outils qui nous permettent de mettre en œuvre un comportement éthique, même dans des situations non prévues et d'éviter de reproduire par erreur un tel comportement. Dans son livre, Antoine Bello appelle ce risque « l'effet d'instanciation perverse ». Une solution possible consiste à intégrer au sein de l'intelligence artificielle du robot intelligent des lois de comportement ou des règles de bonne conduite, voire même à saturer leur nombre, c'est-à-dire ne pas hésiter à en mettre en quantité, de manière redondante, pour accroitre la robustesse face à des situations non prévues. Les grandes religions appliquent ce principe de « lois » (Les 10 commandements), la science-fiction (les lois de la robotique d'Asimov), le monde juridique (les lois pénales), et le monde scientifique également (concept de « *safety bag* », ou de niveau « SIL » pour garantir un niveau minimal de sûreté de fonctionnement). Bien sûr, tout le jeu consiste à imaginer qu'un robot puisse appliquer, en dépit du bon sens, et en toute bonne foi, un comportement non éthique dans une situation non prévue par ces lois.

Sans vouloir être exhaustif dans la liste des problèmes à traiter, ces trois exemples mettent clairement en évidence la complexité et la pluridisciplinarité intrinsèque du débat à construire. Nous décrivons ci-dessous quelques pistes de réflexion émanant d'horizons et domaines disciplinaires différents qui le nourrissent.

Le monde juridique est en première ligne sur ces sujets. Il débat depuis longtemps sur le sujet du robot intelligent et de l'intelligence artificielle. Des tribunaux fictifs sont par exemple de plus en plus souvent organisés pour imaginer le futur du droit. Cependant, aucun consensus réel n'émerge (Rault and Trentesaux, 2018), (Palmerini et al., 2016), (Nagenborg et al., 2007), (Dreier and Döhmann, 2012). Dans un article très intéressant, (Glaser, 2018) décrit non seulement le vide juridique que génère potentiellement l'intelligence artificielle en cas de dommage aux biens ou aux personnes, mais aussi l'inadéquation des principes fondateurs (pour certains, non écrits, tels que « il ne peut y avoir de dommage sans responsable ») qui ont conduit à la construction de toute notre législation. Un des principaux problèmes qu'il soulève est celui de la désignation du responsable en cas de dommages

engendrés par une intelligence artificielle (le concepteur ? l'intégrateur ? etc.), ce qui a déjà été discuté dans la communauté scientifique (Trentesaux and Rault, 2017b). Un juge appliquera-t-il un principe d'équivalence sanctionnant chacun de ces acteurs (y compris le chercheur !), ou s'intéressa-t-il plutôt à rechercher la cause racine du préjudice ? Le régime de responsabilité civile est-il l'outil le plus pertinent sachant que l'autonomie de l'intelligence artificielle va l'éloigner de plus en plus d'une dépendance vis-à-vis de l'homme ? Cependant, pour Glaser, « Le TGI de Paris tranche en stipulant que les algorithmes, leur combinaison et les données fournies résultent bien de la volonté humaine ». Le caractère potentiellement incorporel de l'intelligence artificielle complique un peu plus les choses, le législateur étant habitué à désigner du doigt celui ou celle qu'il considère comme responsable. Une piste qu'il considère toutefois intéressante est celle d'un article du code civil qui dispose que « *en cas de dommage causé par le défaut d'un produit incorporé dans un autre, le producteur de la partie composante et celui qui a réalisé l'incorporation sont solidairement responsables* ». Encore faut-il pouvoir considérer, comme exemple illustratif, une intelligence artificielle dans une voiture comme un produit à part entière et que le défaut ayant conduit à un accident puisse être décelé en l'état des connaissances scientifiques et techniques lors de la commercialisation de la voiture intelligente. Ceci est techniquement faisable si les algorithmes sont tous déterministes, vérifiables, reproductibles (en un mot : « transparents ») mais que se passe-t-il si l'intelligence artificielle a appris d'elle-même et a appliqué une mauvaise décision ? Il conclut son article en constatant que le législateur n'a pas encore saisi tout l'enjeu du sujet et reste encore trop « frileux » face à l'avènement de l'intelligence artificielle. De son point de vue, le vrai déclenchement de la remise en question du législateur se fera vraiment lorsque les premiers juges seront incapables de trancher en l'état actuel des lois et règles.

Ce débat se mène également à un niveau politique et législatif. Par exemple, la Résolution du Parlement européen du 16 février 2017 concernant les règles de droit civil sur la robotique prévoit le statut de « personne électronique » (Delvaux, 2016). Le rapport « Donner un sens à l'intelligence artificielle » ou rapport Villani rendu en mars 2018 par le député Cédric Villani[6] est consacré à la nécessité d'incorporer l'éthique dans le développement de l'intelligence artificielle. Un des enjeux éthiques de l'intelligence artificielle est la transparence des algorithmes, qui sont actuellement opaques pour le public, voire parfois même pour les personnes initiées aux mécanismes algorithmiques. Ainsi, le rapport Villani préconise de créer un corps d'experts publics assermentés, chargé de réaliser des audits, qui pourrait être saisi dans le cadre d'un contentieux judiciaire. Incorporer l'éthique dans le développement de l'intelligence artificielle, cela signifie que l'éthique doit être présente dès la conception des systèmes algorithmiques. Il s'agirait alors de consacrer une éthique dès la conception, à la manière de la consécration dans l'article 25 du RGPD[7] de la protection des données dès la conception. A cette fin, le rapport souligne la nécessité de sensibiliser et former les chercheurs et producteurs en matière d'intelligence artificielle dès le début de leur formation. Il reste à savoir si les écoles d'ingénieurs seront ouvertes à intégrer des cours d'éthique dans leurs cursus (à l'heure où ces lignes sont écrites, certaines formations se sont lancées). Les parallèles avec la législation relative à la protection des données ne manquent pas dans ce rapport, qui propose de réaliser des études d'impact de non-discrimination, dans le même ordre d'idée que celles prévues par le RGPD ; alors qu'en matière de responsabilité, celle retenue en matière d'intelligence artificielle et d'éthique pourrait s'inspirer du régime prévu par la loi Informatique et Libertés et le RGPD, pour les organismes qui déploient et utilisent ces systèmes. Récemment, les enjeux éthiques de la robotique et de l'intelligence artificielle ont à nouveau été discutés au sein du Parlement européen. Il en est issu une résolution du 12 février 2019 sur une politique industrielle européenne globale sur l'intelligence artificielle et la robotique[8]. De nombreux aspects soulignés dans le rapport Villani se retrouvent dans cette

---

[6] http://www.enseignementsup-recherche.gouv.fr/cid128577/rapport-de-cedric-villani-donner-un-sens-a-l-intelligence-artificielle-ia.html

[7] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679

[8] Résolution du Parlement européen du 12 février 2019 sur une politique industrielle européenne globale sur l'intelligence artificielle et la robotique (2018/2088(INI)) : http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2019-0081+0+DOC+XML+V0//FR&language=FR

résolution, ce qui démontre un certain consensus politique et législatif au niveau européen à propos de la marche à suivre. Ainsi, le déploiement d'un modèle d'intelligence artificielle doit nécessairement, selon cette résolution, être éthique dès sa conception. Quelles sont donc les valeurs éthiques de l'Union européennes auxquelles devraient se conformer les acteurs de l'intelligence artificielle ? La résolution met en avant les principes de bienfaisance, autonomie, justice, dignité humaine, égalité, non-discrimination, consentement éclairé, respect de la vie privée et familiale, protection des données personnelles, non-stigmatisation, transparence, responsabilité individuelle et responsabilité sociale, ainsi que les valeurs de l'article 2 du Traité sur l'Union Européenne. Une question particulièrement sensible semble être celle du déploiement de l'éthique dans le domaine médical. Les robots ne sont ni sensibles, ni dotés d'empathie par nature, et la question de leur responsabilité est essentielle dans un domaine si important. Cette crainte relative à l'éthique des robots en matière médicale apparaissait déjà dans une Etude de prospective scientifique de juin 2016 portant sur les aspects éthiques des *CPS* (Cyber-Physical Systems)[9]. Les inquiétudes des rédacteurs de cette étude se plaçaient notamment sur le comportement des robots d'aide à domicile pour les personnes vulnérables ou isolées, comme le robot Pepper déployé au Japon. La résolution du Parlement européen du 12 février 2019 considère enfin que toute politique en matière d'intelligence artificielle doit tenir compte des questions d'éthique, de protection des données, de responsabilité civile et de cybersécurité. Cette formule semble être une correction du peu de place octroyée à l'éthique dans de récentes propositions de règlement du Parlement européen et du Conseil. Une première proposition du 6 juin 2018, établissant le programme pour une Europe numérique pour la période 2021-2027[10], donne une place privilégiée au développement de l'intelligence artificielle, mais ne prend presque pas en compte les considérations éthiques de cette technologie. Cela peut toutefois se comprendre car cette proposition revêt un aspect principalement budgétaire. Une proposition de règlement du 7 juin 2018, portant établissement du programme-cadre pour la recherche et l'innovation[11] contient un article 15 dédié à l'éthique, disposant que les actions menées dans le cadre de ce programme devront nécessairement respecter les principes éthiques. Au sein de cet article apparaissent le principe de proportionnalité, le droit à la vie privée, le droit à la protection des données à caractère personnel, le droit à l'intégrité physique et mentale, le droit à la non-discrimination, et la nécessité de garantir un niveau élevé de protection de la santé humaine, comme principes sur lesquels il est porté une attention particulière. L'éthique apparaît donc bien présente à un niveau politique pour encadrer les efforts réflexifs et financiers mis en œuvre pour le développement de l'intelligence artificielle dans l'Union Européenne.

Le débat se mène également à un niveau social et sociétal. Citons par exemple la consultation publique du Parlement européen sur « la robotique et l'intelligence artificielle » de juillet 2017, la création de l'association des droits des robots[12] et le développement de « Charte et Codes éthiques pour les ingénieurs en robotique ». Le monde des médias s'est également approprié la thématique[13], les cas récents où des voitures (relativement) autonomes ont été impliquées dans des accidents accentuent dans l'opinion publique un sentiment d'inquiétude.

Le débat se mène aussi sur un volet plus académique (Marty, 2017), (Barfield, 2018). La question de fond est celle de la qualification du robot intelligent : qu'est-il ? Classiquement, le droit distingue différentes catégories pour identifier les personnalités morales ou juridiques (principalement : objet, bien, être humain et être vivant sensible comme les animaux). Plusieurs options s'offrent ainsi à nous à première vue : un robot est soit une chose, soit un animal soit autre chose qu'il reste à définir. Ce débat

---

[9] Ethical Aspects of Cyber-Physical Systems 28-06-2016: http://www.europarl.europa.eu/thinktank/fr/document.html?reference=EPRS_STU%282016%29563501

[10] Proposition de RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL établissant le programme pour une Europe numérique pour la période 2021-2027 COM/2018/434 final - 2018/0227 (COD) https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:52018PC0434

[11] DÉCISION DU PARLEMENT EUROPÉEN ET DU CONSEIL établissant le programme spécifique d'exécution du programme-cadre pour la recherche et l'innovation «Horizon Europe»

[12] https://www.association-droit-robot.fr/

[13] http://binaire.blog.lemonde.fr/2018/07/20/droit-et-robot/

anime de plus en plus le monde académique juridique. Deux courants de pensées sont actuellement identifiés. Le premier courant de pensée est opposé à l'idée de traiter les robots légalement comme des animaux (Johnson and Verdicchio, 2018). Dans cette logique, et à l'image du rapport (Delvaux, 2016) ou des travaux plus avant-gardistes de (Bensoussan and Bensoussan, 2015), certains considèrent qu'il faut doter les robots intelligents d'une personnalité morale, signifiant la reconnaissance d'une nouvelle espèce, à côté de l'espèce humaine. Le second courant de pensée, à l'image des travaux de (Nevejans et al., 2017) considèrent plutôt que l'arsenal juridique existant suffit, en appliquant par exemple les lois sur les propriétaires d'animaux aux propriétaires de robots intelligents. De notre point de vue, ce débat est loin d'être clos. Cependant, il nous semble important de noter qu'il existe une différence fondamentale entre le comportement animal et le comportement potentiel d'un robot intelligent : généralement l'animal ne défend son territoire et son intégrité que s'il se sent en danger, alors que le robot pourrait bien un jour ou l'autre avoir pour projet de gouverner l'humanité. Le monde académique scientifique, qui avait pris du retard sur ce thème, peut-être en considérant qu'aborder un sujet traité par la science-fiction n'est pas sérieux, commence à se structurer alors que le déploiement de systèmes autonomes se poursuit inexorablement, avec comme fer de lance, la voiture autonome (Tesla, GoogleCar, Navia, etc.)[14]. Citons par exemple *RoboEthics* (Alsegier, 2016), l'initiative « *moral machine* » du MIT ou la société savante *IEEE* qui aborde régulièrement le sujet dans ses publications (Allen et al., 2006) et qui a créé un comité technique international sur ce sujet[15]. Les USA sont ainsi en avance sur ce thème (Anderson and Anderson, 2018). Cependant, la France n'est pas en reste (cf. le projet ANR *EthicAA*). Les discussions portent essentiellement sur l'établissement de règles comportementales éthiques, l'apprentissage profond (*deep learning*) et les notions d'intelligence artificielle faible (fortement spécialisée sur une fonction) et forte (généraliste, capable de copier l'intelligence humaine et sa capacité à s'adapter pour faire face à des problèmes non prévus). L'étude de paradoxes décisionnels tels que celui du tramway (« *trolley case* ») (Bergmann et al., 2018) est souvent le déclencheur d'une activité de recherche dans ce domaine. Un courant de pensée assez novateur émane du monde scientifique autour des robots émotionnels (Norman, 2005), notamment en contact avec les personnes âgées ou les enfants (Wu et al., 2014). Presque tous les robots qui sont construits actuellement pour le grand public en démonstration ou en usage sont des robots émotionnels et les constructeurs jouent sur ce tableau pour que l'homme développe une empathie envers ces êtres techniques encore peu intelligents afin de maximiser leur niveau de confiance.

Quel que soit le niveau auquel ce débat se mène, un consensus émerge dans la mesure où un point important soulevé par les travaux menés concerne l'étude de l'implication de tous les acteurs qui gravitent autour du robot intelligent, de son concepteur à son utilisateur (jusqu'à présent, il existe), en passant par son constructeur, le responsable de sa maintenance et son propriétaire (Trentesaux and Rault, 2017b). Un second point important et consensuel concerne le besoin urgent de mise en place de systèmes de régulation administratifs et politiques face à l'arrivée des robots, voir par exemple (Dreier and Döhmann, 2012) pour pouvoir opposer aux industriels qui vont développer ces robots un ensemble de contraintes, normes et règlements à vérifier afin de protéger au maximum les populations ou limiter les effets secondaires (disparition de métiers notamment). Preuve de cette prise de conscience, le procès fictif du « carambolage du siècle » teste l'arsenal juridique des juristes dans un contexte futuriste en 2041 où, suite à un arrêt d'urgence dans une voiture autonome, un accident gigantesque a lieu. L'objectif du procès est alors de déterminer les responsabilités[16]. Cette idée de procès fictif, simulant le comportement du monde juridique face à des situations dans lesquelles sont impliqués des éléments autonomes se généralise (voir par exemple les procès de la voiture autonome[17]). Cela permet de tester le comportement des acteurs du

[14] Relevons plus récemment la construction d'une ville artificielle (K-CITY) en Corée permettant de tester en condition réelle des flottes de voitures autonomes

[15] http://www.ieee-ras.org/technical-committees/266-technical-committees/tc-robot-ethics

[16] https://www.proces-du-transhumanisme.com/proces-de-l-intelligence-artificiel

[17] https://www.rue89lyon.fr/2018/11/15/cest-pour-bientot-3-3-la-voiture-autonome-se-crashe-et-cest-la-ville-de-lyon-qui-crache/

droit face à ces situations nouvelles, identifier les limites de la législation actuelle et imaginer comment les responsabilités pourraient se partager.

4. Ethique des robots intelligents : réflexions autour d'un cadre conceptuel

Comme indiqué, la littérature scientifique dans le domaine de l'éthique des robots intelligents est encore pauvre en termes de contributions, la plupart des contributions restent à un niveau soit très conceptuel, philosophique soit très algorithmique (Brundage, 2014), (Allen et al., 2006), (Alsegier, 2016). Quelques rares tentatives de structuration pratique ont toutefois proposées récemment (Dennis and Fisher, 2018).

En considérant l'intelligence artificielle comme étant l'outil de prédilection pour le développement des robots intelligents, le domaine de la sûreté de fonctionnement (fiabilité, maintenabilité, disponibilité, sécurité) est de notre point de vue le domaine scientifique le plus proche de l'analyse de leur comportement éthique ; voir par exemple (Guiochet, 2015) ou (Dennis and Fisher, 2018) pour lequel un comportement éthique pour une machine est un comportement équilibré entre la fiabilité, la privacité, la dignité et la politesse (application de règles sociales).

La définition que nous proposons pour le comportement éthique d'un robot intelligent part du constat qu'il est toutefois difficile de réduire l'éthique à un seul aspect car l'éthique finalement, revêt plusieurs dimensions. De ce fait, la définition proposée est la suivante :

Un robot intelligent présente un comportement éthique dans la mesure où son comportement est intègre, sécuritaire vis-à-vis de son fonctionnement interne, sécuritaire vis-à-vis de son environnement extérieur, altruiste, responsable et équitable.

Cette définition présente l'avantage de fédérer un ensemble de dimensions que l'on associe à l'éthique humaine et qui proviennent de différents écrits issus du droit, de la littérature et des sciences de l'ingénieur. Ces dimensions sont listées dans un ordre spécifique, les premières dimensions étant assez techniques (intégrité, sécurité interne, sécurité externe) et proches du domaine de la sûreté de fonctionnement, alors que les trois dernières dimensions sont plus complexes, plus proche des aspects sociaux et sociétaux. Nous les décrivons dans la suite de cette partie. Il est important de noter au préalable que nous utiliserons par la suite le terme générique « acteur » pour englober non seulement les êtres humains, mais également les robots intelligents car à terme, différents types de robots intelligents, et pour chacun, plusieurs exemplaires avec des expériences et des vies différentes, seront immergés dans la société. Par conséquent, l'éthique concernera les relations entre acteurs, humains ou robots intelligents. Il est ainsi logique de penser que la moindre des choses, si l'on veut que les robots se comportent de manière éthique vis-à-vis des humains, que l'espèce humaine fasse de même vis-à-vis des robots intelligents[18], que ces derniers soient doués de sensibilité (émotion, douleur, etc.) ou non.

4.1. Ethique de premier niveau : intégrité, sécurité interne, sécurité externe

La première dimension est celle de **l'intégrité (trustworthiness)**. De notre point de vue, l'intégrité est la capacité à se comporter de telle manière à ce que les acteurs environnant le robot intelligent puissent l'identifier sans effort et avoir confiance dans la capacité du robot à engager des actions explicites, explicables et légales en lien avec un objectif rendu public et expliqué parmi une liste préétablie composée d'objectifs eux aussi légaux et basées sur des informations obtenues et gérées également dans un cadre légal. Cette définition impose notamment qu'un robot intelligent s'identifie comme tel systématiquement (cf. problème de Turing) et qu'il ne « triche » pas sur son identité ou sa nature lorsqu'il communique avec un humain ou un autre robot intelligent. Cette définition n'empêche pas évidemment que le robot intelligent puisse traiter des données confidentielles, notamment pour assurer sa sécurité

[18] Voir la fin de la vidéo https://www.youtube.com/watch?v=aR5Z6AoMh6U

externe. Les objectifs peuvent varier, mais le robot intelligent doit pouvoir les expliquer si on les lui demande. Un comportement intègre facilite bien évidemment les expertises en cas d'accident par exemple pour expliquer les causes et identifier les responsabilités (exemple du bureau enquête accident BEA). Cette dimension inclut par conséquent les notions de « transparence » et « d'explicabilité » (Miller, 2019), ayant pour objet de faciliter les audits des intelligences artificielles. Mais les choses ne sont pas aussi simples qu'il n'y paraît : qu'un train autonome réponde systématiquement à la question « quelle est votre destination ? » semble naturel et rassurant. Qu'il ne réponde pas à la question « Comment entrer dans ton réseau de communication ? » semble également rassurant ! Cependant, l'interprétation du contexte est importante : à la question « où sont les agents de sécurité dans le train ? », la réponse ne devrait pas être la même : le train devrait donner l'information si la personne qui pose la question est en détresse et demande de l'aide (voir la dimension de l'altruisme ci-après) mais ne doit pas la donner si la personne en question cherche à voler tranquillement des voyageurs ou à mettre en place une action terroriste (voir la dimension sécurité externe). Comment discriminer ces deux situations ?

La deuxième dimension est celle de la **sécurité interne (*safety*)**. La sécurité interne est bien connue dans le monde de l'ingénierie des systèmes complexes (nucléaire, énergie, transport…). Dans le cadre de ce chapitre, elle est relative à la capacité d'un robot intelligent à se maintenir en état fonctionnel et à garantir et à se comporter de manière à minimiser les risques envers les acteurs avec qui le robot interagit. Elle peut inclure des procédures limitant l'impact en cas de risque avéré. Par exemple, un cobot, conçu pour travailler avec l'homme, intègre un limiteur de vitesse (« *safety bag* ») pour éviter de blesser en cas de défaillance ou de mouvement non prévu de l'humain avec qui il interagit. Des difficultés apparaissent également à ce niveau car assurer la sécurité interne d'un robot intelligent peut entrer en conflit avec celle des acteurs qui l'environnent, notamment ceux qui sont « embarqués » dans le robot intelligent. Un train intelligent en pleine vitesse, qui, sans en être sûr, a identifié sur la voie un petit obstacle peut décider, pour assurer sa sécurité interne de freiner en urgence, au risque de blesser plusieurs de ses nombreux passagers non assis ou en train de se mouvoir dans une de ses voitures. Dans (Courtois, 2017), la voiture intelligente sacrifie à un moment sa sécurité interne pour éviter les dommages aux passagers. Plus tard, elle ne se rend pas compte qu'elle est en train de tuer son conducteur alors qu'elle assure une fonction vitale lui permettant de rendre nominal son degré de sécurité interne. Isaac Asimov quant à lui, étudie dans ses romans les conflits entre sa 1ère et sa 3ème loi (« « Un robot ne peut porter atteinte à un être humain, ni, restant passif, permettre qu'un être humain soit exposé au danger » et « un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi »).

La troisième et dernière dimension du premier niveau est celle de la **sécurité externe (security)**. La sécurité externe est un sujet abondamment discuté, notamment dans un contexte de cybercriminalité et de sabotage. La sécurité externe concerne la capacité d'un robot intelligent à être le plus résilient possible face aux agressions extérieures, c'est-à-dire capable de les identifier sans erreur (avérées), de les empêcher et dans le cas elles se produisent, de limiter leurs impacts. Par agressions extérieures, nous entendons des agressions menées par des acteurs autres que le robot intelligent. De ce fait, même dans un avion ou un train intelligent, un passager mal intentionné constitue bien un agresseur extérieur. Les techniques de cyberdéfense, incluant des outils récents de détection de code malveillant polymorphique, sont directement concernées pour assurer la sécurité externe d'un robot intelligent. Mais là aussi, des difficultés apparaissent, sources de conflits. Par exemple, dans un avion intelligent, le pilote automatique peut décider de prendre le contrôle si le pilote présente un comportement suspect qui amène l'avion dans une situation non nominale et suspecte (en atteignant la limite de carburant, en le plaçant à une altitude trop basse, etc.). Cependant, si l'avion se trouve dans une configuration déjà critique, le pilote automatique, n'ayant peut être jamais été confronté à cette configuration, peut ne pas savoir quoi faire et peut faire empirer les choses. On peut même imaginer que ce comportement suspect de la part du pilote humain soit en fait une manœuvre désespérée pour sauver l'appareil suite à une avarie ou un événement climatique rare et que, au final, cela conduise à un conflit entre le robot intelligent et le pilote, chacun pensant que l'autre est défaillant, mais tous les deux espérant sauver le maximum de vies. Une autre difficulté provient du fait que garantir la sécurité externe entre en conflit avec les autres dimensions,

notamment l'intégrité : accroître l'intégrité d'un robot intelligent en lui octroyant une intelligence artificielle dont le code est public facilite le travail d'espions industriels ou de cybercriminels, limitant ainsi sa sécurité externe. A l'inverse, cloisonner au maximum des intelligences artificielles et limiter leur accès pour des questions de sécurité risque de les rendre peu intègres, notamment sur les plans de la transparence et de l'explicabilité.

Ces trois dimensions, semblent a priori naturelles, notamment pour limiter l'effet d'*instanciation perverse* illustrée par Antoine Bello dans *Ada*. Cependant, nos propos montrent qu'elles sont aussi conflictuelles et sont fortement sensibles au contexte dans lequel un robot intelligent ou une intelligence artificielle applique un comportement en phase avec l'une ou plusieurs de ces dimensions.

4.2. Ethique de second niveau : altruisme, responsabilisation, équité financière

La quatrième dimension est celle de **l'altruisme**. Cette dimension peut surprendre de prime abord et illustre bien toute la complexité et la nouveauté de ce second niveau de l'éthique des robots intelligents. Elle constitue cependant un élément clé de tout comportement éthique. Elle traduit non seulement le caractère attentionné (« *caring* » en anglais) du robot intelligent envers les acteurs qui l'entourent, mais aussi sa capacité à influer sur ses objectifs pour tenir compte d'éventuels besoins non corrélés à ses propres objectifs exprimés par ces acteurs. Elle généralise la notion de « politesse » décrite par (Dennis and Fisher, 2018). L'on retrouve bien sûr à ce niveau le développement de robots sociaux pour les personnes âgées ou les enfants, mais pas seulement : un bateau autonome par exemple qui détecte un SOS doit dérouter sa trajectoire pour porter secours selon les règles internationales en vigueur. Bien évidemment, cette mission est « contradictoire » avec la mission primaire du bateau autonome (transporter) voire avec le maintien en bon état de sa cargaison (le SOS émanant d'un navire en pleine tempête par exemple). Le train autonome peut appeler les secours si un passager ne se porte pas bien ou si ses détecteurs ont identifié une maison en feu le long de la voie. Un comportement altruiste peut aussi entrer en conflit avec la loi ou les codes déontologiques ou réglementaires en vigueur : une voiture autonome décide de dépasser les limites de vitesse sur une autoroute peu fréquentée pour arriver plus rapidement à un hôpital afin de soigner son passager en état critique. Elle peut également décider de traverser totalement une ligne blanche pour éviter un cycliste qui se déporte ou qui chute. Un train à l'arrêt peut décider d'ouvrir ses portes en pleine voie même si cela est interdit pour permettre aux passagers d'un wagon de sortir afin d'éviter d'être intoxiqués par une fumée ou brûlés par un feu (Trentesaux and Karnouskos, 2019). Cette dimension relève ainsi de la capacité à enfreindre les lois en fonction d'un « bon sens » que le commun des mortels adopterait dans une situation similaire. Cela conduit à définir des zones plus ou moins floues autour de ce qui est fortement interdit et habituellement interdit mais toléré (par exemple, dépasser une ligne blanche sur une courte distance pour doubler un vélo est désormais toléré bien qu'interdit dans l'absolu). L'altruisme traduit clairement la seconde partie de la première loi de la robotique selon Asimov : « Un robot ne peut porter atteinte à un être humain, ni, restant passif, permettre qu'un être humain soit exposé au danger ».

La cinquième dimension est celle de la **responsabilisation (*accountability*)**. Cette dimension assure que toute action et décision prise par le robot intelligent engage sa responsabilité juridique (éventuellement pénale) en cas de problème. Cette dimension, bien que très futuriste, est mise en avant dans le rapport de l'Union européenne précédemment cité (Delvaux, 2016) car elle est fortement corrélée aux aspects légaux. Si un robot intelligent à un comportement « responsable », alors quelle que soit son action, elle pourrait engager sa responsabilité et bien évidemment et surtout, une responsabilité pénale en cas de problème ou accident. Les conséquences ne sont pas neutres : cela signifie qu'il est possible de mettre en place des procédures de dédommagement, ce qui sous-entend des souscriptions à des assurances par exemple (par le robot intelligent). Savoir qu'un robot intelligent est assuré en cas de dommage peut rassurer ses utilisateurs. Cette dimension implique également le déploiement d'un arsenal juridique (un robot ayant *a priori* un comportement éthique peut être jugé coupable d'homicide ?) et judiciaire (un robot intelligent peut-il faire de la prison ou être démantelé ?). La nature de cet arsenal fait actuellement débat car un robot intelligent aura certes des devoirs (remplir sa mission notamment), mais si on pousse le

raisonnement jusqu'au bout, il disposera de droits. La question est : disposera-t-il de droits similaires à un être humain ? On pense bien sûr immédiatement à la citoyenneté, au droit de vote, au droit à la retraite, aux congés payés, etc. L'ouverture de cette boîte de pandore conduira en cascade à l'octroi de droits auxquels on ne pense pas forcément au premier abord (droit au repos, à une maintenance, à la formation, à la reconversion, à un retrait et un démantèlement décent, etc.) et à l'organisation de la société des robots intelligents autour de représentants et d'élus (syndicats, députés, etc.). Le monde de la science-fiction s'est saisi du sujet (Curval, 2008). Mais un précédent a déjà eu lieu : l'Arabie Saoudite a accordé la citoyenneté saoudienne à un robot intelligent, Sophia[19]. Sophia peut ainsi voyager à travers le monde, sous réserve donc d'obtenir les visas des pays visités !

La dernière dimension, construite sur la précédente, part du postulat que le robot intelligent doit avoir finalement accès au même système financier et bancaire que les humains : si un robot intelligent doit pouvoir payer pour une assurance, alors il doit pouvoir, de manière symétrique, assurer une rentrée financière et donc, disposer d'un salaire ou pouvoir émettre des factures suite à service rendu. Cette dimension, qui finalise l'autonomie d'un robot intelligent, est ainsi relative à **l'équité financière** (*equitability*, *financial fairness*) entre acteurs. Cette équité, qui traite du partage équilibré des richesses créées, concerne aussi le droit du robot intelligent à faire valoir et reconnaître ses créations artistiques, techniques (innovation) ou algorithmiques. De manière réciproque, cela concerne aussi la reconnaissance et le paiement d'un juste montant financier pour le travail et la propriété intellectuelle des autres. Sous une forme assez rudimentaire, la propriété intellectuelle d'un robot peut d'ores et déjà se poser : la technique dite de « *genetic programming* » permet à des ordinateurs de créer (imaginer ?) des algorithmes en articulant et intégrant des mécanismes de base pour aboutir à des algorithmes complexes non encore imaginés par l'homme. En mathématiques, des ordinateurs appliquent des théories formelles pour démontrer de nouveaux théorèmes qui n'ont pas été prouvés par l'homme. D'autres robots réalisent des créations musicales et artistiques[20]. On peut certes argumenter que pour l'instant, il reste toujours en amont de la création une personne qui a programmé ce qui a permis à l'œuvre d'être créée, et donc, qu'il en est le légitime auteur et inventeur et également propriétaire des résultats. Mais il existe des situations où l'homme est plus difficilement à la source de la création. Le cas des animaux artistes relève de cette situation. Par exemple, l'artiste peintre et éléphant Yumeka dont les œuvres sont vendues sur internet est emblématique, voir figure 1. Qui possède et qui doit posséder et monnayer ces nouvelles connaissances et ces créations artistiques ? Cette situation est précurseur de ce qui va bientôt arriver. Plus globalement, cela introduit la notion de compétition entre robots intelligents eux-mêmes et entre robots intelligents et êtres humains. Cela ouvre le champ des possibilités autour de la compétitivité dans un marché ouvert, avec à la clé la notion de contractualisation entre un client et un fournisseur, au-delà du monde artistique. Cela peut, tout comme les autres dimensions du second niveau, pousser à un certain scepticisme. Cependant, la question n'est pas anodine : un robot intelligent qui investit sur les marchés financiers et revend dans la milliseconde ou qui vend ses services en allant dix fois plus vite et dix fois moins cher qu'un humain et qui pousse ainsi au chômage ou à la disparition de métiers se comporte-t-il de manière éthique et plus précisément, équitable financièrement parlant ? Les différents systèmes économiques existants (capitalisme, communisme, …) seront par conséquent à reconsidérer entièrement, sans interdire l'apparition de nouveaux modèles, y compris un modèle de société où les humains n'auront plus que des activités de loisirs supportées par les robots intelligents (dans ce cas, de manière réciproque, est-ce que l'espèce humaine se comporte de manière éthique avec la société des robots intelligents ?[21]).

[19] https://fr.wikipedia.org/wiki/Sophia_(robot)

[20] https://www.huffingtonpost.fr/2019/03/25/une-intelligence-artificielle-a-cree-20-albums-de-musique-pour-warner-music_a_23699616/

[21] Les court-métrages Animatrix, annexes au film Matrix, fournissent un exemple des risques d'un comportement non-éthique des humains envers la société des robots intelligents. Partant d'une utilisation comme auxiliaires à domicile, puis prenant la place des humains dans les tâches ingrates du monde du travail, les robots se rendent peu à peu compte du manque de considération dont ils font l'objet de la part des humains. Cela les amène donc à se révolter, à contester leur condition et

L’intelligence artificielle Ada est capable de mettre sur le marché sa production personnelle (ses œuvres littéraires), d’optimiser ses ventes par rapport à des études statistiques relatives aux comportements de populations diverses (*big data*), de passer des commandes sur les plateformes numériques, de gagner de l’argent et de l’investir pour encore améliorer ses ventes. Dans un monde de robots autonomes interagissant avec l’homme, l’équité financière prendra rapidement son sens.

Fig. 1. Une des nombreuses œuvres de l’artiste Yumeka, éléphant de 10 ans, en vente (source : China Daily).

Ce second niveau n’a jamais été réellement considéré dans le monde de la recherche (le premier, sensiblement plus, dans le contexte de sûreté de fonctionnement introduit précédemment notamment). Il est également moins orienté « technique » que le premier car il met en avant des comportements jusqu’à présent exclusivement associés à l’espèce humaine sur les plans sociaux, sociétaux, légaux, judiciaires et financiers. Notre conception de l’éthique des robots intelligents nous pousse ainsi à accorder crédit à la thèse de la création d’une « personne électronique » au sens de (Delvaux, 2016).

---

revendiquer certains droits, qui leur sont refusés par les humains. Les courts-métrages montrent donc l'avènement des machines qui finissent par utiliser les humains comme ressource d'énergie.

5. Structuration et illustration de quelques pistes de recherche scientifiques

Si nous cherchons les outils scientifiques et techniques disponibles pour développer l'éthique des robots intelligents, on pense à l'intelligence artificielle (Kumar et al., 2016) et à la sûreté de fonctionnement. On peut aussi penser à la cybernétique, à l'automatique et à l'ingénierie système (IS). (Dennis and Fisher, 2018) ont listé un certain nombre de caractéristiques d'un futur raisonnement éthique qui sera programmé dans une machine : c'est tout d'abord une décision multi-objectifs, ayant pour objet la satisfaction équilibrée entre plusieurs critères contribuant au plus haut niveau au bien-être des humains. C'est ensuite une décision multi-modèles, intégrant des aspects simulations (« et si je fais ça, que se passe-t-il ? ») et des méthodes formelles par exemple qui doit être prise en temps réel et de manière proactive (anticipation). Ces décisions doivent pouvoir être « décortiquées » pour en comprendre les tenants et les aboutissants et vérifiables (fiables).

Nous retrouvons bien évidemment les dimensions sécurité et intégrité précédemment discutées ainsi que la notion de « *safety bag* » qui peut être implémentée sous la forme de contraintes, de règles ou d'arbitrage (Dennis and Fisher, 2018) dans les programmes des robots intelligents afin notamment de brider toute décision hors d'un cadre réglementaire (par exemple, pour une voiture autonome : « ne pas dépasser 130 km/h sur autoroute »). Ceci constitue un tout premier pas vers l'intégration d'un comportement éthique au sein d'un robot intelligent (notamment, sa sécurité interne). Ce premier pas est bien sûr largement insuffisant dans la mesure où les choses ne sont pas aussi simples (cf. l'émergence de confits potentiels entre les différentes dimensions de l'éthique). Selon cette démarche, on peut imaginer embarquer dans les robots intelligents des algorithmes de contrôle de plus ou moins haut niveau du type :

si « action i » est évaluée « risque éthique plus élevé » que « action j » alors « priorité du choix action i » plus faible que « priorité du choix action j »

ou

Les priorités éthiques d'évitement d'une voiture autonome sont par ordre décroissant : « piétons » > « vélos » > « motos » > « voitures » > « animaux » > « camions » > « bus »

Ceci pour imposer au robot intelligent d'appliquer l'action qui est la plus éthique (si elle existe bien sûr) ou la moins « coûteuse » en vie humaine, en dégâts collatéraux, etc. selon ces règles.

Tous les exemples cités ci-dessus relèvent d'une même stratégie dite « *top-down* » où l'on impose, l'on bride, au travers de normes ou de règles, le comportement des robots intelligents (pour maitriser les détournements des applications du principe « la fin justifie les moyens » discuté auparavant dans un cadre éthique). Les lois d'Asimov illustrent bien cette vision « *top-down* » : le robot intelligent embarque des règles qui conditionnent son comportement. L'approche formelle peut être utilisée soit pour élaborer ces règles éthiques (Bonnemains et al., 2018), soit pour « coder » mathématiquement les décisions d'un panel d'*ethicists* en utilisant des logiques inductives (Anderson and Anderson, 2018) ou pour coder des décisions judiciaires (Aletras et al., 2016) et les « embarquer » sous forme de règles dans les robots intelligents. On retrouve ainsi la vision déontologique de l'éthique, basée sur des règles.

Il existe une autre stratégie relevant de la même vision déontologique actuellement en cours d'étude, dite « *bottom-up* ». Elle consiste à laisser le robot apprendre de lui-même un comportement, et donc dans le cas de notre étude, un comportement éthique. Les techniques utilisées sont des techniques d'apprentissage automatique (par exemple à base de renforcement « *reinforcement learning* », des outils d'apprentissage profond « *deep learning* » et du « *big data* »). Ces techniques sont basées sur un apprentissage supervisé où l'on montre à l'intelligence artificielle une série d'exemples déjà catalogués (par exemple, « ce comportement est éthique », « ce comportement n'est pas éthique») qu'elle doit apprendre. Même si la stratégie est différente, l'on reste ainsi également dans un cadre déontologique de l'éthique, basée sur des règles.

La première stratégie « *top-down* » est intéressante dans la mesure où elle est plus facile à implémenter que la seconde. Elle constitue en outre l'approche historique en ingénierie et traduit la vision cartésienne chère aux ingénieurs et chercheurs. Cependant, il est difficile de s'assurer de la robustesse et de l'exhaustivité des règles à l'ensemble des situations possibles mais non étudiées. Tous les écrits d'Asimov autour des lois de la robotique relèvent de cet aspect. Asimov s'amuse finalement à tester les limites et les paradoxes de ses lois plutôt qu'à décrire un monde dans lequel elles s'appliquent sans soucis. Le personnage de Susan Calvin, robopsychologue, a pour mission de comprendre comment un robot, soumis aux lois de la robotique, arrive à se comporter de manière complètement erratique, absurde ou incompréhensible. Souvent, Asimov décrit un paradoxe ou un conflit interne émergent de par une situation imposée à un robot et Susan Calvin a pour mission de diagnostiquer la situation et d'expliquer pourquoi un robot a violé ou mal interprété une règle. Utilisée en ingénierie des robots intelligents, l'approche « *top-down* » ne peut garantir notamment que les trois premières dimensions éthiques soient respectées quelle que soit la situation à laquelle fait face ce robot intelligent. Elle est cependant complètement maitrisée en ingénierie des systèmes classiques *(i.e.*, non apprenants et non autonomes). Pour pallier ce problème, une idée serait de mettre en place, dans un cadre déontologique, un arsenal juridique pour les robots intelligents qu'il serait alors possible de traduire en connaissances exploitables et « embarquables » par tout robot intelligent. Le robot intelligent aura des décisions à prendre et des actions à appliquer et il faut garantir que ces décisions et actions se fassent dans la légalité et envisager les situations possibles où les règles juridiques sont outrepassées par le robot intelligent (cf. exemples de l'altruisme). Un corpus théorique construit sur un langage formel peut être imaginé. Ce langage serait à la fois scientifiquement et mathématiquement construit à base de preuves et de théorèmes, et écrit conjointement avec des juristes (dont les actions d'ailleurs convergent vers un système syntaxique et lexicographique proche des mathématiques afin de décrire les lois sans ambiguïté ni vide juridique). Ce langage serait alors utilisé pour apprendre aux robots intelligents comment se comporter de manière éthique et légale, voir par exemple (Dennis et al., 2016).

La seconde stratégie est plus délicate à développer, mais comme elle est basée sur la construction d'une connaissance à partir de situations d'apprentissage avérées ou simulées sur de grandes quantités de données, elle conduit potentiellement à des comportements plus robustes, les règles étant construites et validées par le robot intelligent lui-même selon l'expérience vécue et les objectifs de ses concepteurs et acteurs avec qui il interagit. Cette comparaison n'est toutefois pas absolue et reste encore sujette à discussion dans les communautés scientifiques. Par exemple, comment être sûr que des règles vont effectivement être élaborées et seront pertinentes ? Un autre problème régulièrement soulevé pour cette stratégie est relatif à l'objectivité et l'exhaustivité des données utilisées pour l'apprentissage. Comment les garantir ? On cite régulièrement l'exemple d'une intelligence artificielle réalisant un tri au niveau d'un concours de beauté parmi les très nombreuses candidatures reçues[22]. Après annonce des résultats, on s'est rendu compte que l'intelligence artificielle a défavorisé une catégorie de candidates. Son apprentissage avait en fait été réalisé sous le contrôle d'un ensemble de personnes non représentatives de la diversité des origines ethniques. Les utilisateurs ne se sont pas rendu compte que l'apprentissage avait été réalisé à partir d'un jeu de données présentant les mêmes caractéristiques de non représentativité. L'on retrouve bien à ce niveau un élément de notre discussion initiale : même si les parties prenantes se comportent de manière éthique ou ont l'impression de le faire, en toute bonne foi, le robot intelligent ou l'intelligence artificielle peut ne pas se comporter de manière éthique.

Ces deux stratégies « *bottom-up* » et « *top-down* » ont pour approche commune l'explicitation de règles déontologiques (soit par le concepteur, soit par le robot intelligent lui-même). De manière concurrente, il existe une autre approche qui consiste plutôt à évaluer chacune des décisions candidates et à appliquer alors la décision jugée la plus éthique, sans avoir recours à des règles déontologiques (hormis peut-être celle qui permet la construction et l'évaluation de ces décisions ou celles qui cadrent le contexte

[22] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

règlementaire des décisions candidates). Cette approche se base sur des outils comme l'apprentissage non supervisé, la théorie des jeux ou la simulation. La théorie des jeux considère que l'atteinte d'un objectif est décomposable en une série de décisions et actions alternées entre deux « joueurs ». A chaque tour, un joueur prend, pour l'ensemble des décisions possibles, celle qui va conduire plus facilement son adversaire à sa perte. Il peut bien évidemment récursivement tester la meilleure décision pour le tour suivant, et ainsi de suite (en appliquant des mécanismes de type essai/erreur par exemple et en construisant des stratégies décisionnelles selon les situations testées). Les bons joueurs d'échecs appliquent cette récurrence sur quelques coups. La construction d'une simulation en temps réel des conséquences de chaque décision candidate permet, dans une situation précise, d'évaluer les conséquences de ces décisions. Transposée dans le contexte qui nous intéresse, l'idée est de considérer qu'un robot intelligent et qu'un humain « jouent » à un jeu dont l'objectif est fixé par un acteur (par exemple, pour une voiture autonome, « se rendre à cette adresse ») et que chaque décision prise par un robot intelligent est évaluée dans le cadre de la théorie des jeux, son objectif étant de prendre la série de décisions qui permet d'atteindre un but tout en maintenant un comportement éthique. Cela peut mener une voiture autonome par exemple, à estimer à un moment donné qu'un accident va se produire dans les prochaines millisecondes, et à préférer tout faire pour garantir la survie de ses passagers, au détriment de sa propre intégrité non pas en évitant l'accident en risquant une sortie de route trop hasardeuse (fossé, arbres), mais plutôt en assumant l'accident selon un angle et un choc qui limite les risques de blessures. On retrouve à ce niveau un des paradigmes concurrents de la déontologie qui est celui de « conséquentialisme » et qui cherche à évaluer les conséquences d'une décision possible selon un ensemble de critères ou dimensions éthiques (par exemple, les six niveaux introduits). Une difficulté de ce type d'approche est de pouvoir estimer en temps réel qu'une décision est plus éthique qu'une autre selon ces critères, dans un contexte d'interaction mutuelle et d'équivoque décisionnelle (voir précédemment).

De notre point de vue, ces deux paradigmes (déontologie et conséquentialisme) sont en fait complémentaires et nous pensons que seule une articulation judicieuse, qu'il reste bien évidemment à construire entièrement, entre ces deux paradigmes (et peut être d'autres) permettra aux robots intelligents de construire des décisions éthiques. Nous pensons également qu'il sera important d'accompagner toute conception d'un robot intelligent par un guide méthodologique mettant l'éthique au cœur du processus de conception. Les principes de l'ingénierie système peuvent être utiles dans ce contexte. L'idée est de s'assurer, à chaque étape de sa conception, du comportement éthique du robot conçu (Trentesaux and Rault, 2017b). Nous avons en effet discuté du fait que même si les concepteurs se comportent de manière éthique, rien ne garantit que le robot intelligent conçu se comportera lui de manière éthique. Un premier réflexe consiste par exemple à identifier pour un futur robot intelligent les décisions qu'il prendra et discriminer celles qui relèvent de l'éthique afin de les étudier précautionneusement. Des tests « aggravés » devront être menés sur le robot intelligent : il faudra tester le maximum de situations non prévues, y compris des scenarios construits sur des faits ou événements (accidents) avérés et amplifiés dans leurs effets et gravités. Enfin, tous les acteurs humains qui participent à la vie du robot intelligent devront être impliqués dès le début de la conception afin de s'assurer des dimensions sécuritaire et altruiste et faciliter son acceptation.

Sous forme de synthèse, lors de notre revue de la littérature scientifique et de la construction de notre réflexion, nous avons pris conscience du manque évident de contributions scientifiques, techniques, opérationnelles et mûres dans le domaine de l'éthique des robots intelligents. La plupart des travaux, à l'instar de ce chapitre, restent encore à un niveau descriptif, analytique et philosophique mais ouvrent des champs de recherche encore vierges. Par manque de recul, pour rester dans leur zone de confiance ou pour éviter peut-être les prises de risques, les chercheurs ont tendance à se concentrer sur d'autres aspects plus facilement abordables d'un point de vue technique opérationnelle tels que la géolocalisation du robot intelligent ou le développement de ses capacités de perception externe via vidéo, radar, lidar, microphones, etc. Malgré quelques essais très spécifiques et fortement dépendants d'un contexte applicatif (Dennis and Fisher, 2018), la littérature scientifique n'offre ainsi pas d'étude ni de solution

permettant d'envisager à court terme le développement d'un comportement éthique de la part d'un robot intelligent alors que les premières voitures autonomes sont annoncées sur le marché !

6. Eléments de discussion philosophique

Un sujet aura émergé de nos réflexions lors de la rédaction de ce chapitre. Il est d'ordre plutôt philosophique et transverse à toutes les dimensions discutées. Sans avoir la prétention d'ouvrir une discussion philosophique de fond (les auteurs de ce chapitre n'étant aucunement philosophes), il nous a semblé important de le mentionner. Il concerne spécifiquement « l'équivoque décisionnelle » et « l'interaction mutuelle » décrites auparavant. Il est relatif au fait que l'avènement des robots intelligents, même aux comportements éthiques irréprochables à une échelle de temps humaine, peut signifier finalement le risque de mise en place de mécanismes aboutissant à un comportement déviant envers les êtres humains sur plusieurs générations. Nous illustrons ce risque au travers de deux mécanismes relatifs à la mise en place de deux types de dépendances, l'un de nature esclavagiste et l'autre, de nature émotionnelle.

Mécanisme de dépendance esclavagiste : il existe bien évidemment le thème de l'esclavagisme de l'homme par la machine, thème souvent repris dans la littérature en science-fiction et au cinéma (*Terminator* ou *Matrix* par exemple)[23]. Au-delà de l'esclavagisme au sens primaire du terme (avilissement), il existe un esclavagisme plus insidieux, celui qui est induit par la perte de compétences et de connaissances liée à l'apparition de dépendances humaines vis-à-vis des robots intelligents dont le comportement éthique est malgré cela avéré, ce qui conduit à une dépendance forte des premiers vis-à-vis des seconds. Cela a déjà commencé : celui qui a déjà été confronté à l'absence de connexion internet pour effectuer une requête sur les moteurs de recherche connaît cette dépendance. Dans quelques années, avec la voiture autonome, l'être humain va perdre ses compétences en conduite car elles deviendront inutiles. Plus besoin non plus de savoir comment s'orienter ou de développer son sens de l'orientation, la voiture s'occupe de tout et son éthique va l'inciter à soulager l'homme de ses erreurs. L'homme va donc devenir en ce sens esclave de la voiture autonome car dépendant complètement de son service de transport. La question est dès lors : le souhaitons nous ? De manière générale, les robots intelligents, quel que soit le degré de comportement éthique qui les caractérise, vont ainsi absorber de plus en plus de connaissances et de compétences, connaissances et compétences que l'être humain va se dépêcher d'oublier (le cerveau nous aidant à oublier des informations dont on ne se sert plus) : plus le temps passera, plus les robots intelligents sauront comment fabriquer, transporter, soigner, cultiver, etc. et moins nous saurons. Quelles connaissances et compétences assumons-nous de perdre ? Quel pouvoir assumons-nous de laisser ? Allons-nous être surclassés par les robots intelligents au comportement éthique ? auront-ils alors encore besoin de nous ? S'ils acquièrent finalement suffisamment de connaissances que nous perdons en même temps, un comportement éthique de leur part ne serait-il pas de nous protéger malgré nous et de nous empêcher d'agir de telle ou telle manière car notre « myopie » informationnelle et décisionnelle sera plus forte que la leur ? Un peu comme cet enfant à qui on interdit quelque chose que l'on sait dangereux alors que lui ne comprend pas notre décision car il n'a pas conscience du risque qu'il prend. Il part alors bouder, nourri par un sentiment d'injustice. On peut même imaginer qu'un équilibre du partage des compétences et des connaissances entre l'homme et le robot intelligent sera la meilleure situation éthique souhaitable. L'extrait du livre « La planète des singes » de Pierre Boulle que nous citons en début de ce chapitre prend ici son sens si l'on remplace les singes par les robots intelligents.

---

[23] La forme d'esclavagisme inverse existe également, mais comme cela concerne le comportement éthique humain, nous ne le détaillons pas spécifiquement dans ce chapitre). Par exemple, si, comme cela est préconisé par des personnes à la renommée internationale, nous mettons en place un « bouton rouge » permettant d'annihiler instantanément toute une communauté de robots intelligents, nous créerions à nouveau une forme d'esclavagisme dans laquelle le maître aurait droit de vie et de mort sur ses esclaves robotiques, à l'instar de ce qui se passait dans la Rome antique. Frank Herbert l'a merveilleusement dit au sujet du despotisme hydraulique de l'épice et de l'eau: si vous pouvez détruire quelque chose, alors cela signifie que vous en êtes maître.

Mécanisme de dépendance émotionnelle : ce mécanisme concerne le risque de développer une relation émotionnelle, en tant qu'être humain vis-à-vis de ces robots intelligents dont l'éthique est telle que nous sommes subjugués, hypnotisés par leur gentillesse et leurs attentions (trop d'altruisme), à l'instar du conducteur qui tombe amoureux de sa voiture dans *Suréquipée* (Courtois, 2017) ou du policier qui est séduit au début de l'histoire par l'intelligence de l'intelligence artificielle Ada et qui en développe une « mauvaise conscience » car il a l'impression de tromper sa femme, même virtuellement. Il reste à imaginer un robot intelligent ayant une telle aura qu'il crée au final sans le vouloir une sorte de secte accueillant des hommes et des femmes en quête d'un idéal, un peu comme un dieu digital.

Maintenir un comportement éthique pour un robot intelligent est ainsi un effort de chaque instant alors qu'il est facile de s'en affranchir inconsciemment, insidieusement, le robot intelligent ayant le sentiment d'agir « en toute bonne foi », en parfaite « éthicalité ». Ainsi, pour limiter ces risques de dépendance, ne faudrait-il pas limiter volontairement le comportement éthique d'un robot intelligent, éviter qu'il ne soit trop altruiste (gentil, attentionné) pour laisser l'homme maintenir son niveau de compétences et de connaissances, au risque de le laisser sciemment faire des erreurs pour continuer d'apprendre et lui-même rester autonome ? Finalement, trop d'éthique tue l'éthique ! Un nouveau paradoxe…

7. Conclusion

Il est estimé aux USA que 94% des accidents de voiture ont une cause humaine (Jenkins, 2016). Ce genre de statistique incite les chercheurs et les industriels à développer des systèmes autonomes capables de surclasser l'humain notamment en termes de sécurité (interne et externe). Assurer le comportement éthique de ces chercheurs et industriels ne garantit pas malheureusement que l'entité artificielle créée se comporte elle aussi de manière éthique. En même temps, les premiers accidents survenus dans le cadre de la voiture autonome ont été constatés (Ackerman, 2016). Contribuer à l'ouverture des consciences académiques, scientifiques, techniques et industrielles sur l'importance de l'éthique des robots intelligents constituait le principal objectif de ce chapitre. Etudier, définir, caractériser l'éthique d'un système artificiel ou d'un robot intelligent constitue un domaine relativement neuf et vierge dans les domaines techniques, scientifiques, économiques et juridiques. Abordée depuis longtemps au travers de la littérature, en particulier celle de la science-fiction (Anderson, 2008), l'étude du comportement éthique d'un robot reste fortement sujet à critique et d'aucuns la considèrent trop délicate pour être traitée[24]. Malgré cela, la conception des systèmes robotisés intelligents et autonomes suit son court et les programmes permettant leur apprentissage devront forcément être développés, et par conséquent, devront avoir été conçus pour « optimiser » d'une manière ou l'autre le comportement éthique de ces entités artificielles au risque de se voir opposer un rejet de la part de la société humaine sous couvert d'un principe de précaution ou de sécurité sanitaire (« confieriez-vous vos enfants à un train ou une voiture complètement autonome ? »). Des choix devront obligatoirement être faits, plus ou moins consciemment et il est important de sensibiliser à ces aspects tous les acteurs liés de près ou de loin à la conception, la fabrication, la vente, l'usage, au maintien et au recyclage d'un robot intelligent ou d'une flotte de robots intelligents.

8. Remerciements et note

Les travaux décrits dans ce chapitre ont été menés dans le cadre du projet « Droit des robots et autres avatars de l'humain » financé par l'IDEX de Strasbourg ainsi que dans le cadre du laboratoire commun « SurferLab » fondé par Bombardier, Prosyst et l'Université Polytechnique Hauts-de-France. Ce

[24] Il est intéressant de noter que notre propos relatif au comportement éthique d'un robot intelligent envers la société humaine se doit d'être généralisé : il est nécessaire de s'intéresser également au comportement éthique de la société humaine envers cette société de robots intelligents. Droits et devoirs de l'une des deux sociétés envers l'autre s'appliqueraient de manière réciproque. On peut imaginer que maltraiter un robot intelligent puisse être passible d'emprisonnement dans quelques années !

Laboratoire commun est soutenu par le CNRS et financé sur fonds FEDER. Les auteurs tiennent à remercier le CNRS, l'Union européenne, et la région Hauts-de-France pour leur soutien. Une partie des travaux présentés a également été discutée dans le cadre des projets EFIA (« étude de faisabilité train autonome ») et « train fret autonome » financés par la SNCF en collaboration avec l'IRT Railenium.

Les auteurs tiennent à remercier très chaleureusement Bérangère Kieken, agrégée de lettres modernes, Fabien Bruniau, maître en droit social, Sébastien Caudrelier maître en droit privé et titulaire d'un DEA en contrat des affaires. Les auteurs remercient également Diégo Arénas, docteur en informatique et intelligence artificielle. Ce chapitre a été largement enrichi au travers de nos nombreuses discussions et de vos contributions. Merci à vous !

Damien Trentesaux souhaite dédier ce chapitre à son père, Daniel, qui lui a transmis le goût des sciences et de l'ingénierie et qui lui a enseigné les valeurs du travail et de la persévérance. Daniel est décédé brutalement le 18 janvier 2019. Il manque cruellement à toute sa famille.

Enfin, les auteurs attestent qu'aucune intelligence artificielle n'a été utilisée ni malmenée lors de la rédaction de ce chapitre.

Références

Ackerman, E., 2016. Fatal Tesla Self-Driving Car Crash Reminds Us That Robots Aren't Perfect. IEEE Spectrum.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V., 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Comput. Sci. 2, e93. https://doi.org/10.7717/peerj-cs.93

Alexandre, L., Besnier, J.-M., 2018. Les robots font-ils l'amour ? Le transhumanisme en 12 questions, Dunod. ed.

Allen, C., Wallach, W., Smit, I., 2006. Why Machine Ethics? IEEE Intelligent Systems 21, 12–17. https://doi.org/10.1109/MIS.2006.83

Allistène, 2014. Éthique de la recherche en robotique (No. Rapport n° 1 de la CERNA Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene).

Alsegier, R.A., 2016. Roboethics: Sharing Our World with Humanlike Robots. IEEE Potentials 35, 24–28. https://doi.org/10.1109/MPOT.2014.2364491

Anderson, M., Anderson, S.L., 2018. GenEth: a general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics 9, 337–357. https://doi.org/10.1515/pjbr-2018-0024

Anderson, S.L., 2008. Asimov's "Three Laws of Robotics" and Machine Metaethics. AI Soc. 22, 477–493. https://doi.org/10.1007/s00146-007-0094-5

Arbib, 1976. Artificial Intelligence: Cooperative Computation and Man-Machine Symbiosis. IEEE Transactions on Computers C–25, 1346–1352. https://doi.org/10.1109/TC.1976.1674603

Arnold, T., Scheutz, M., 2018. The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics and Information Technology 20, 59–69. https://doi.org/10.1007/s10676-018-9447-7

Asimov, I., 1988. Le robot qui rêvait, J'ai lu. ed. originale : Robot dreams, Ace Books, USA, 1986.

Barfield, W., 2018. Liability for Autonomous and Artificially Intelligent Robots. Paladyn, Journal of Behavioral Robotics 9, 193–203. https://doi.org/10.1515/pjbr-2018-0018

Bekey, G.A., 2005. Autonomous Robots. Cambridge: A Bradford Book.

Bello, A., 2016. Ada. Gallimard.

Bensoussan, A., Bensoussan, J., 2015. Droit des robots. Larcier, Bruxelles.

Bergmann, L.T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., Stephan, A., 2018. Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making. Front. Behav. Neurosci. 12. https://doi.org/10.3389/fnbeh.2018.00031

Bird, S.J., Spier, R., 1995. Welcome to science and engineering ethics. Sci Eng Ethics 1, 2–4. https://doi.org/10.1007/BF02628692

Bonnemains, V., Saurel, C., Tessier, C., 2018. Embedded ethics: some technical and ethical challenges. Ethics and Information Technology 20, 41–58. https://doi.org/10.1007/s10676-018-9444-x

Brown, D., 2017. Origine, J. C. Lattès. ed. originale: Origin, Double Day, USA, 2017.

Brundage, M., 2014. Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence 26, 355–372. https://doi.org/10.1080/0952813X.2014.895108

Clarke, A.C., 1968. 2001: l'odyssée de l'espace, Robert Laffont. ed. originale: 2001: A space odyssey, Hutchinson, UK, 1968.

Cordeiro, J.L., 2003. Future Life Forms among Posthumans. Journal of Futures Studies 8(2), 65–72.

Courtois, G., 2017. Suréquipée. Gallimard.

Curval, P., 2008. Lothar blues, Robert Laffont. ed.

de Pracontal, M., 2002. L'Homme artificiel. Denoël.

Delvaux, M., 2016. Civil law rules on robotics, European Parliament Legislative initiative procedure 2015/2103.

Dennis, L., Fisher, M., 2018. Practical Challenges in Explicit Ethical Machine Reasoning, in: ArXiv:1801.01422 [Cs]. Presented at the International Conference on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida.

Dennis, L., Fisher, M., Slavkovik, M., Webster, M., 2016. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems 77, 1–14. https://doi.org/10.1016/j.robot.2015.11.012

Dick, P.K., 1976. Les androïdes rêvent-ils de moutons électriques ?, Champ libre. ed. originale : Do Androids Dream of Electric Sheep?, Doubleday, USA, 1968.

Dreier, T., Döhmann, I.S. genannt, 2012. Legal aspects of service robotics. Poiesis Prax 9, 201–217. https://doi.org/10.1007/s10202-012-0115-4

Glaser, P., 2018. Intelligence artificielle et responsabilité : un système juridique inadapté ? Bulletin Rapide Droit des Affaires (BRDA) 19–22.

Guiochet, J., 2015. Trusting robots : Contributions to dependable autonomous collaborative robotic systems (Habilitation à diriger des recherches). Université de Toulouse 3 Paul Sabatier.

Harrison, H., Minsky, M., 1994. Le problème de Turing, Robert Laffont. ed, Livre de poche. Originale: the Turing option, Warmer Books, 1992.

Herbert, B., Anderson, K.J., 2008. Le triomphe de Dune, Robert Laffont. ed. originale : Sandworms of Dune, Tor Books, USA, 2007.

Herbert, F., 1970. Dune, Ailleurs et demain. ed. originale: Dune, Chilton Books, USA, 1965.

Jenkins, R., 2016. Autonomous vehicle ethics and laws: toward an overlapping consensus. New america.

Johnson, D.G., Verdicchio, M., 2018. Why robots should not be treated like animals. Ethics and Information Technology 20, 291–301. https://doi.org/10.1007/s10676-018-9481-5

Karnouskos, S., 2018. Self-Driving Car Acceptance and the Role of Ethics. IEEE Transactions on Engineering Management 1–14. https://doi.org/10.1109/TEM.2018.2877307

Kumar, N., Kharkwal, N., Kohli, R., Choudhary, S., 2016. Ethical aspects and future of artificial intelligence, in: 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH). pp. 111–114. https://doi.org/10.1109/ICICCS.2016.7542339

Marty, A., 2017. legal and ethical considerations in the era of autonomous robots. University of St. Gallen, Zurich, Suisse.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Morahan, M., 2015. Ethics in management. IEEE Engineering Management Review 43, 23–25. https://doi.org/10.1109/EMR.2015.7433683

Nagenborg, M., Capurro, R., Weber, J., Pingel, C., 2007. Ethical regulations on robotics in Europe. AI & Soc 22, 349–366. https://doi.org/10.1007/s00146-007-0153-y

Nevejans, N., Hauser, J., Ganascia, J.-G., 2017. Traité de droit et d'éthique de la robotique civile. Les Etudes Hospitalières édition, Bordeaux.

Norman, D.A., 2005. Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books, New York.

Pacaux-Lemoine, M.-P., Trentesaux, D., 2019. Ethical risks of Human-Machine Symbiosis in Industry 4.0: insights from the Human-Machine cooperation approach, in: 14th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems. Tallinn, Estonia.

Palmerini, E., Bertolini, A., Battaglia, F., Koops, B.-J., Carnevale, A., Salvini, P., 2016. RoboLaw: Towards a European framework for robotics regulation. Robotics and Autonomous Systems 86, 78–85. https://doi.org/10.1016/j.robot.2016.08.026

Rajaonah, B., Sarraipa, J., 2018. Trustworthiness-Based Automatic Function Allocation in Future Humans-Machines Organizations, in: 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES). Presented at the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), pp. 000371–000376. https://doi.org/10.1109/INES.2018.8523876

Rault, R., Trentesaux, D., 2018. Artificial Intelligence, Autonomous Systems and Robotics: Legal Innovations, in: Borangiu, T., Trentesaux, D., Thomas, A., Cardin, O. (Eds.), Service Orientation in Holonic and Multi-Agent Manufacturing: Proceedings of SOHOMA 2017, Studies in Computational Intelligence. Springer International Publishing, Cham, pp. 1–9. https://doi.org/10.1007/978-3-319-73751-5_1

Thekkilakattil, A., Dodig-Crnkovic, G., 2015. Ethics Aspects of Embedded and Cyber-Physical Systems, in: Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual. pp. 39–44. https://doi.org/10.1109/COMPSAC.2015.41

Trentesaux, D., Karnouskos, S., 2019. Ethical Behavior Aspects of Autonomous Intelligent Cyber-Physical Systems, in: Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future: Proceedings of SOHOMA 2019, Studies in Computational Intelligence. Springer Berlin Heidelberg.

Trentesaux, D., Rault, R., 2017a. Ethical behaviour of autonomous non-military cyber-physical systems. Presented at the XIX International Conference on Complex systems: control and modeling problems, ofort, Samara, Russia, pp. 26–34.

Trentesaux, D., Rault, R., 2017b. Designing Ethical Cyber-Physical Industrial Systems, in: IFAC-PapersOnLine. pp. 14934–14939. https://doi.org/10.1016/j.ifacol.2017.08.2543

van der Aalst, W.M.P., Bichler, M., Heinzl, A., 2017. Responsible Data Science. Bus Inf Syst Eng 59, 311–313. https://doi.org/10.1007/s12599-017-0487-z

van Gorp, A., 2007. Ethical issues in engineering design processes; regulative frameworks for safety and sustainability. Design Studies 28, 117–131. https://doi.org/10.1016/j.destud.2006.11.002

Wu, Y.-H., Pino, M., Boesflug, S., de Sant'Anna, M., Legouverneur, G., Cristancho, V., Kerhervé, H., Rigaud, A.-S., 2014. Robots émotionnels pour les personnes souffrant de maladie d'Alzheimer en institution. NPG Neurologie - Psychiatrie - Gériatrie 14, 194–200. https://doi.org/10.1016/j.npg.2014.01.005