The Implementation of Ethical Decision Procedures in Autonomous Systems : the Case of the Autonomous Vehicle

Katherine Evans

SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE V

UMR 8224 Sciences, Normes, Démocratie

T H È S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ

Discipline : Philosophie

Présentée et soutenue par :

Katherine EVANS

le 15 janvier 2021

The Implementation of Ethical Decision Procedures in Autonomous Systems

The Case of the Autonomous Vehicle

Sous la direction de :

M. Stéphane CHAUVIER – Professeur, Sorbonne Université

Membres du jury :

Mme Joanna BRYSON – Professeur, Hertie School of Governance
M. Raja CHATILA – Professeur, Sorbonne Université
M. Christoph LÜTGE – Professeur, Technical University of Munich
M. John SULLINS – Professeur, Sonoma State University

Acknowledgements

It is a popular truism to say that it takes a village to raise a child, but it is my experience that it takes the population of multiple continents to write a thesis. The work concealed in the pages that follow has been the product of what seems at times like a global conversation, one that has spanned across 4 years, and given me many experiences I never could have hoped for. Accordingly, I have many people to thank for this journey.

Firstly, I owe a great deal to the members of the AVEthics project, for their encouragement, their expertise, and their enthusiasm in regard to my unconventional ideas. Specifically, I would like to thank Ebru Dogan for the opportunity to work on this project, and for teaching me what it means to work in an interdisciplinary team. To this end, I feel a deep gratitude for my partner in arms, Nelson de Moura, who beyond providing emotional support throughout this thesis, also provided the much-needed computational support for every one of its ideas. To Raja Chatila, I owe both the confidence and clarity with which I approach machine ethics, as well as my very material existence in France throughout a portion of this thesis. Finally, I would like to thank my supervisor, Stéphane Chauvier, for first believing in the project of this thesis, and throughout the years, for the rare gift of having taught me how to think from my own point of view, mostly by concealing his.

I am also deeply grateful to the people whom I have met throughout my professional experiences within this thesis. I am grateful to the many friends I have at the Vedecom Institute for welcoming a philosopher into their midst. I am also incredibly lucky to have encountered the great mind of Geoff Keeling throughout my work in autonomous vehicle ethics, without whom many of the ideas of this thesis would have been left undiscovered. I owe much to Noah Goodall for allowing me to ‘put the passenger first’ in front of a room full of AV industry leaders, and for welcoming me into the Anglo-Saxon world of autonomous vehicles. Finally, I owe an enormous debt to Sasha Rubel for trusting me enough to let me talk on many UNESCO panels, and through her particular social prowess, for having led me towards incredibly fruitful ideas with Louisa Zanoun, Nicolas Miailhe, Peter Burgess, and Marc-Antoine Dilhac.

Finally, I am incredibly grateful for the extensive cast of characters I call family, both by blood and by circumstance. I would like to thank Captain Alan and the staff and patrons of Bar Chez Camille, who unwittingly participated in a number of informal acceptability studies on autonomous vehicles, and who wittingly accepted the presence of a true *philosophe du comptoir*. I must also give credit to the eight characters of comedy that have lit up my life throughout the darker days of this thesis, and who I can truly call my friends: Cyane Talamoni, Penelope Berger, Marta Russoli, Tristan Thomas, Dayan Oualid, Romain Wibaux, Alice Maniable and Alyssa Brannlund. I would also like to extend a special thank-you to my very first philosophy professor and friend, Andreï Poama, whose casual mentioning of sex robots sparked my interest in machine ethics over 9 years ago. Finally, I would like to thank my parents, Dan and Debra Evans, for having provided every material and emotional comfort required for the accomplishment of this thesis, and who have surreptitiously learned as much as I have about machine ethics, for having proofread every single page I have ever written. I can only hope that the work exposed in the pages that follow does justice to all of your efforts.

This research has been conducted as part of the AVEthics Project funded by the French National Agency for Research, grant agreement number ANR-16-C22-008, and the Vedecom Institute.

Table of Contents

Résumé Substantiel en Français

De nos jours, le monde est peuplé de deux sortes d'agents distincts : humain et *artificiel*. Ensemble, ils œuvrent pour l'accomplissement d'une gamme immense de fins humaines : la régulation de la bourse internationale, la gestion des achats sur des marchés virtuels, le soin des patients et des personnes âgées, voire même la précision létale d'une opération militaire. Il est donc vrai de dire que ces deux sortes d'agents entretiennent un rapport *collaboratif* au sein des sphères sociales humaines variées. Reste que ce serait aller trop vite en besogne que de supposer que la qualité de cette collaboration soit bien définie, ou que les conséquences qui en résultent soient complètement anodines aux yeux de l'éthique.

En effet, parmi cette gamme immense d'agents artificiels, il en existe certains qui opèrent dans des contextes de saillance éthique. Cette saillance est moins fonction du pur *impact* que ces agents auront en vertu de leur existence dans un contexte—on pense ici aux objections éthiques liées à l'installation de mines sur un territoire, ou un système de surveillance en ville— que de la *prise de décision* de ces derniers, et du pouvoir qui en résulte : une voiture autonome, lorsqu'elle est face à une collision imminente et inévitable, doit ostensiblement *choisir* qui survivra entre le passager ou le piéton. Il en va de même pour les robots médicaux qui doivent pondérer le respect de l'autonomie du patient contre la bonne application des recommandations médicales, ou bien les

systèmes d'assistance décisionnelles qui doivent prédire le taux de récidive potentiel d'un individu, ou l'aptitude d'un candidat pour un poste au sein d'une entreprise. En chacun de ces cas, les décisions elles-mêmes que prendront ces machines courent le risque d'avoir un impact liberticide ou nocif, à minima pour les individus qui sont touchés par ces choix, et potentiellement aussi pour la société au sens large.

C'est donc à cette troublante capacité décisionnelle qu'est due l'émergence d'un type spécial d'agent artificiel, l'agent *moral* artificiel. D'un coté, on distingue ces agents sur un plan *contextuel*, puisque l'accomplissement de leurs buts pratiques au sein de certaines sphères sociales peut naturellement courir le risque de nuire aux agents humains, comme le cas du véhicule autonome, ou des armes autonomes. D'un autre coté, on peut distinguer ces agents sur un plan *comportemental* : si un agent humain doit (ou devrait) faire usage de son raisonnement moral en sa capacité de médecin, aide-soignant, recruteur ou assistant personnel, il semble plausible que toute machine qui remplace l'humain dans un tel rôle doive être dotée d'une capacité similaire. Un agent moral artificiel est donc l'agent qu'il semble nécessaire de construire dès lors que l'on cherche à automatiser des taches ou des occupations qui sont sujets de ce mariage éthiquement saillant entre contexte et comportement, faisant en sorte que des contraintes éthiques, ou ce que l'on pourrait appeler une *moralité artificielle*, soient implémentées directement dans le raisonnement de la machine.

A partir de là, il peut sembler naturellement propice de s'inspirer de la moralité *humaine* dans la construction d'une moralité artificielle. Mais cette idée se heurte cependant à deux obstacles aussi bien ontologiques qu'éthiques. En premier lieu, il existe une dissonance problématique et inévitable entre l'ontologie humaine d'un côté, et l'ontologie d'un agent moral artificiel, de l'autre. Tandis que l'humain est typiquement vu comme une créature dotée d'une subjectivité, ou d'une moralité individuelle, autonome, et universelle, le tout laissant place à l'attribution d'un statut d'*agent moral* ; ces mêmes capacités chez l'agent artificiel demeurent relativement peu profondes voire même inexistantes, ce qui perturbe considérablement le bon fonctionnement de concepts essentiels à la moralité comme la responsabilité, l'autonomie, ou l'*akrasia* morale. Il en résulte que toute tentative de simuler la moralité humaine chez la machine sera nécessairement superficielle, syntactique et incomplète. En second lieu, du fait que les

machines n'ont pas un caractère moral inné ou essentiel, il semble suivre qu'il dépend de *nous*, les agents humains, de leur en attribuer un. Reste que la forme ultime que prendrait ce caractère moral demeure mystérieuse dans beaucoup de contextes d'implémentation possibles, faute d'un consensus clair et résolu concernant les normes, valeurs ou principes qui devraient y jouer un rôle clef, ou faute d'une connaissance complète de la vérité morale objective.

En réponse, parmi ceux qui néanmoins cherchent à implémenter une moralité artificielle chez les agents artificiels, deux positions distinctes se sont formées, que l'on peut identifier comme celle des *maximalistes* et celle des *minimalistes.* Au fond, la position du maximaliste relève d'une dépendance des théories morales classiques, et revient à insister sur le fait qu'une bonne moralité artificielle consiste en un programme computationnel qui a) d'un point de vue interne, reproduit parfaitement le processus de raisonnement indiqué par la théorie en question, ou qui b) d'un point de vue externe, semble mener à un comportement décisionnel qui obéit totalement aux recommandations de la théorie en question. En ce sens, l'idéal du maximalisme est l'apparition des machines qui agissent comme de purs maximiseurs d'utilité dans leur environnement, ou qui se comportent comme des agents Kantiens, Rawlsiens, ou Smithiens parfaits. De l'autre côté, la position du minimaliste revient à insister sur le fait qu'une bonne moralité artificielle demande de doter la machine d'une connaissance vaste a) du comportement descriptif des agents humains dans des contextes moraux, ou b) des préférences ou des attitudes morales d'une population donnée. En ce sens, l'idéal du minimalisme est l'apparition des machines qui se comportent selon la volonté ou la sagesse de la foule, ou si cette sagesse ne livre pas de verdicts unis, selon un certain compromis entre ces préférences divergentes.

Néanmoins, force est de constater que ces deux positions restent imparfaites. D'un côté, on peut facilement accuser le maximaliste d'une forme de favoritisme arbitraire pour sa théorie morale préférée, puisque son choix de construire des robots Kantiens plutôt que des robots Rawlsiens ne semble se baser ni sur le caractère inné ou même préféré d'un robot, ni sur la vérité objective des recommandations de la théorie elle-même. De plus, il semble qu'une telle approche mènera nécessairement à une moralité artificielle qui ne s'empare que de certains couleurs de la gamme totale de moralité humaine, et ce faisant, risque d'ignorer des valeurs ou principes que certains êtres humains tiennent pour vrais, ou du moins, qui figurent dans les attentes normatives

de ces derniers. De l'autre côté, l'approche du minimaliste, qui a pourtant plus de chance à satisfaire ces attentes normatives, est néanmoins limitée par la qualité morale de son support empirique. Dans ce sens, quoique l'approche du minimaliste ait la vertu de *plaire* à la foule à travers sa moralité artificielle, elle court le risque d'y intégrer les préférences immorales, prudentielles, biaisées ou bornées qui sont cachées derrière les attitudes de cette dernière.

Un certain chevauchement entre les meilleurs aspects de ces deux approches peut donc sembler nécessaire, afin d'assurer que la moralité artificielle d'un agent moral artificiel reste sensible aux attentes des agents humains, sans pour autant tomber ni dans l'approbation arbitraire d'une théorie morale stérile et inapplicable, ni dans la mobilisation d'un rapport fallacieux entre le comportement moral descriptif de l'humain et le comportement moral idéal d'une machine.

Le but de cette thèse est de creuser une telle approche, en visant l'acceptabilité publique d'un agent moral artificiel comme point de repère.
Nos arguments se diviseront en deux grandes parties : une première partie (chapitres I à III) qui traite les questions ontologiques liées à la construction des agents moraux artificiels, et qui vise à établir avec clarté cette dissonance entre la condition humaine et celle de la machine ; puis une deuxième partie qui traite explicitement de la construction d'une moralité artificielle (chapitres IV à VI) visant à évaluer le genre de contraintes qu'imposent l'ontologie robotique et l'acceptabilité publique à cet égard. Le chapitre VI et VII présenteront notre réponse aux problèmes des minimalistes et maximalistes, sous forme d'une théorie de moralité artificielle, la *théorie des valences éthiques*. Cette présentation prendra une forme plus générale en chapitre VI, mais sera approfondie à travers le cas des véhicules autonomes au chapitre VII.

Chapitre I : *Agent et environnement*

Les agents artificiels constituent une catégorie immense d'entités dans le monde. En effet, la seule caractéristique qui les regroupe catégoriquement est ce que l'on peut appeler leur *ontologie télique* [purpose-oriented ontology] : le fait qu'un agent artificiel est construit par un être humain afin d'accomplir des taches ou des activités précises dans son environnement. Or ce qui contribue

à leur diversité catégoriale est précisément cette diversité télique, et la grande variété d'environnements qui peuvent servir comme un locus d'implémentation : le monde réel, le monde virtuel, un hôpital, les routes parisiennes, les achats d'un individu, les courses de personnes âgées, les candidats à une embauche, etc.

De plus, la notion d'automation est elle-même susceptible de degrés de complexité différents : un thermostat peut détecter des changements environnementaux et y réagir, mais il n'est pas pour autant évident de les assimiler à des machines industrielles capables de construire une maison entière sans la moindre intervention humaine. Il en résulte qu'il est difficile de fournir une définition exhaustive et claire d'un agent artificiel. Néanmoins, notre analyse nous fera aboutir à la définition suivante :

> Un agent artificiel est un artefact technologique doté d'une capacité d'action flexible au sein d'un[1] *Umwelt* particulier, afin d'accomplir ses buts concrets. Cette action flexible comprend : (1) une capacité *réactive*, qui permet à la fois la détection des aspects saillants de son environnement, une responsivité aux changements qui peuvent y survenir, mais aussi une capacité d'adaptation ou d'apprentissage suite à ces changements, mobilisant ce que l'agent sait du monde qui l'entoure [world knowledge]. (2) une capacité *proactive* qui permet la poursuite de ses buts internes en fonction des aspects saillants de son *Umwelt*. (3) une capacité sociale qui permet la communication des buts internes de l'agent aux entités présentes dans son *Umwelt*, ainsi que la capacité d'appréhender les buts et les intentions de ces entités.

Si notre définition semble particulière, elle l'est surement grâce à son usage du mot « Umwelt ». Ce concept, conçu pour décrire le rapport entre un animal et son environnement[2], désigne ici un environnement pratique restreint dans lequel l'agent artificiel est contraint d'agir. En ce sens, l'ontologie télique d'un agent artificiel correspond aux détails de son *Umwelt* : une voiture autonome doit être équipée pour détecter des aspects saillants comme les piétons ou les indications de route, et non pas pour distinguer entre un combattant et un civil. Un soldat robotique, en revanche, doit être doté d'une capacité proactive qui assure l'accomplissement de sa mission militaire, en fonction des membres de son bataillon, et non pas en fonction de la volonté d'un piéton distrait. Parler de l'*Umwelt* d'un agent artificiel souligne donc sa spécificité ontologique, comportementale et télique.

[1] Le mot *Umwelt* est féminin en allemand, mais « monde » étant masculin en français, nous écrirons *un* Umwelt.
[2] Firenze, 2019: von Uexküll, 1957: Heidegger, 1995.

Malgré la spécificité individuelle de chaque sorte d'agent artificiel, il existe néanmoins quatre catégories assez larges auxquelles un agent artificiel peut appartenir. Ces catégories sont relatives aux aspects ontologiques de l'agent, notamment le type de programme computationnel qui le fait passer de la perception à l'action dans un environnement, mais aussi les conditions matérielles de cet environnement même. En ce sens, un agent artificiel peut être *incarné,* doté d'une existence physique (et souvent d'une mobilité) dans le monde matériel, ou bien *virtuel*, opérant uniquement dans un espace virtuel comme l'internet ou des réseaux sociaux. De plus, un agent artificiel peut être doté d'un programme *déterministe*, signifiant qu'un système *explicite* de règles, instructions ou axiomes a été conçu par un être humain et implémenté dans l'agent, ou bien un agent peut être doté d'un programme *stochastique,* ce qui signifie typiquement que l'agent apprend le comportement idéal de manière *tacite,* à travers un processus d'apprentissage profond porté par des réseaux neuronaux artificiels, ou un apprentissage par renforcement. Dans tous les cas, l'agent vise à maximiser sa *mesure de performance* [performance measure] à travers son action et expérience, un critère qui juge l'efficacité de l'agent quant à l'accomplissement de ses buts pratiques : plus l'agent accomplit son télos particulier, plus il fait preuve de rationalité[3].

Tandis que d'un point de vue ontologique, un agent artificiel, et éventuellement un agent *moral* artificiel peut appartenir à n'importe laquelle de ces catégories, il est en revanche plus délicat de dire que le choix entre ces options est complètement anodin d'un point de vue moral. En effet, si la notion d'une moralité artificielle s'applique principalement aux inquiétudes quant à l'impact décisionnel d'un agent moral artificiel, il existe aussi une autre sorte d'inquiétude, qui concerne ce que l'on peut appeler les *problèmes éthiques de conception* [ethical design concerns], qui s'adressent quant à eux à la relation entre les choix de configuration ontologique d'une machine d'un côté, et l'impact de ces choix sur le bien-être humain, de l'autre. Ces problèmes de conception existent principalement sous forme de *principe* de conception normatif[4], visant par exemple la maximisation de la *transparence* des décisions prises par la machine, ou faisant en sorte qu'un agent humain soit toujours *responsable* de ces décisions[5]. En outre, il est important de souligner

[3] Russel & Norvig, 2013, 37.
[4] Jobin et al., 2019.
[5] Dignum, 2019.

que la satisfaction de ces principes n'est pas garantie par la seule présence d'une moralité artificielle : un agent moral artificiel peut donc être moral d'un point de vue décisionnel, mais problématique d'un point de vue de conception.

Dans notre analyse, nous avons choisi de n'explorer que trois principes relevant de ces problèmes éthiques de conception, afin d'en exposer les jalons majeurs : les principes de transparence, de responsabilité et de redevabilité [accountability]. Brièvement, le principe de redevabilité vise la capacité qu'a une machine d'expliquer ses actions et décisions aux utilisateurs. Il est donc un principe technocentrique qui touche la capacité sociale d'un agent artificiel. On pourra facilement imaginer le but et la portée d'un tel principe à travers l'idée d'un bouton qui, une fois pressé, force l'agent artificiel à justifier l'action qu'il vient d'entreprendre[6]. Le principe de responsabilité quant à lui, désigne le rôle que joue l'*humain* dans la conception des agents artificiels, et vise l'établissement de liens explicites dans la chaine décisionnelle qui mène du constructeur vers l'agent artificiel lui-même. En ce sens, adhérer au principe de responsabilité c'est souvent rendre *explicite* et *publique* la personne ou les personnes qui serviront comme locus de responsabilité légale, voire même morale, pour les actions d'une machine. Enfin, le principe de transparence, à son tour, vise l'intelligibilité de la machine elle-même, assurant que les mécanismes d'analyse et de prise de décision sont compréhensibles voire reproductibles par les êtres humains. Il est donc quelque part vrai que la satisfaction des principes de responsabilité et de redevabilité présuppose la satisfaction du principe de transparence.

À la lumière de ces précisions, il semble assez évident que certaines catégories d'agents artificiels passeront plus facilement le cap que d'autres. Principalement, les agents dotés d'un programme déterministe, en vertu du simple fait qu'un être humain ait exhaustivement codé le comportement de cet agent, pourra facilement satisfaire au moins au principe de transparence. Les agents stochastiques, en revanche, souffrent de ce qu'on appelle souvent le *problème de la boite noire* [black box problem], ce qui signifie que leurs décisions, voire mêmes leurs critères et paramètres décisionnels, sont souvent opaques et non reproductibles par les êtres humains[7]. Ce problème est aggravé par le fait qu'une telle opacité laisse ouverte la possibilité que des paramètres

[6] IEEE Global Initiative, 2016, 20: Dignum, 2019, 53.
[7] Eilam, 2005: Dignum, 2019: Castelvecchi, 2016: Oh, Scheile & Fritz, 2019: Gabriel, 2020

éthiquement contestables—comme l'idée d'estimer le récidivisme potentiel d'un délinquant selon ses origines ethniques—opèrent tacitement dans la prise de décision de la machine[8]. De plus, du fait que les agents stochastiques satisfont difficilement le principe de transparence, il semble suivre que leur degré de conformité aux principes de responsabilité et de redevabilité soit aussi remis en question.

Nous n'avons pas choisi d'exposer cette relation entre les choix de programme et les problèmes éthiques de conception afin d'immuniser l'approche déterministe contre toute inquiétude morale, ne serait-ce que parce qu'une telle approche elle-même pose des problèmes techniques que seule une approche stochastique peut facilement résoudre. Ce qu'il nous importe plutôt de suggérer ici, c'est simplement que le choix de l'ontologie de tout agent artificiel passe par un filtre contextuel, mais aussi un filtre éthique, et que ces deux contraintes auront un impact assez fort sur les formes possibles que prendra toute moralité artificielle.

Chapitre II : *l'autonomie et les agents moraux artificiels*

Parmi toutes les capacités qu'il semble nécessaire d'implémenter chez les agents moraux artificiels, il est certain que la notion d'*autonomie* demeure la plus complexe et la plus mal comprise. D'un côté cette confusion vient d'une homonymie entre le concept philosophique d'autonomie (morale) et l'autonomie robotique; mais d'un autre côté, il semble que la notion d'autonomie, en raison de sa signification pour la philosophie morale, met en jeu un autre problème éthique lié à la conception des agents moraux artificiels : celui qui vise à décider si les agents moraux artificiels peuvent, pourront, ou devraient être considérés comme des agents moraux au sens *humain* du terme. En effet, tandis que l'entreprise générale de doter un agent artificiel d'une moralité artificielle semble obéir à une volonté de *mimétisme* avec la moralité humaine, le concept d'autonomie lui-même semble nous pousser à décider *jusqu'où* cette approximation a légitimement lieu d'être.

[8] Dressel & Farid, 2018: Angwin et al., 2016.

Afin d'exposer ces complexités, nous commençons par examiner le concept robotique d'autonomie. En général, l'autonomie au sens robotique souligne l'idée d'une absence de supervision ou de contrôle direct sur le comportement d'une machine. En ce sens, plus une machine peut agir sans l'aide d'un être humain, plus elle est considérée comme autonome. Néanmoins, cette absence de contrôle peut prendre deux sens distincts : un premier sens, que l'on peut qualifier d'*autonomie par provision,* implique qu'un être humain ait fourni toutes les informations ou mécanismes nécessaires à l'action autonome dans un *Umwelt* particulier. C'est-à-dire qu'un agent humain—le programmeur—a prévu toutes les situations rares ou surprenantes qu'une machine pourra rencontrer dans la poursuite de ses buts pratiques. Cette vision de l'autonomie s'aligne donc sur l'approche déterministe, exposée dans le chapitre précédent. En revanche, ce que l'on peut qualifier d'*autonomie par indépendance,* relève cette fois d'une approche plutôt stochastique, qui mise sur la capacité d'apprentissage de l'agent afin qu'il s'ajuste aux changements environnementaux sans l'aide d'un humain.

Avec ces deux distinctions, nous traitons ensuite de ce que l'on peut considérer comme l'*échelle d'autonomie des agents artificiels,* qui admet six niveaux différents. Cette échelle est en quelque sorte une organisation linéaire des distinctions posées entre des variétés d'agents moraux artificiels possibles par des auteurs différents. D'une extrémité à l'autre, cette échelle passe de ce que James Moor appelle des *agents d'impact éthiques*[9], des agents artificiels minimaux qui ne sont pas dotés d'une moralité artificielle et qui sont plutôt sujets d'une moralité d'usage, jusqu'aux agents superintelligents[10], des agents fictifs susceptibles d'appréhender la vérité morale par eux-mêmes et à qui, par conséquent, les êtres humains pourront devoir une sorte de *déférence rationnelle* quant aux décisions morales. Ces deux extrêmes servent à mettre en lumière les vecteurs principaux de cette échelle : la diminution du contrôle de l'être humain sur la machine, et l'approximation, voire éventuellement le dépassement même de la condition humaine.

Tandis que chaque niveau présente ses vertus analytiques, la plupart de nos arguments se concentrent sur le quatrième niveau d'autonomie (celui de ce que l'on peut appeler *des agents éthiques explicites*) et sur sa frontière avec le cinquième niveau, qui est vraisemblablement la place

[9] Moor, 2006.
[10] Bostrom, 2014.

qu'occuperait un agent humain sur cette échelle. D'un point de vue purement descriptif, le statut d'agent éthique explicite comprend la conception standard d'un agent moral artificiel : cet agent ne possède pas la gamme complète des caractéristiques qui semblent nécessaires à la satisfaction des conditions d'un statut d'agent moral humain (conscience de soi, subjectivité, autonomie morale, etc.), mais doit néanmoins être doté d'un système de normes explicites (une moralité artificielle) afin d'assurer la qualité éthique de sa prise de décision. La demande pour cette moralité artificielle, à son tour, est fonction de ce que l'on peut appeler l'*argument de l'automation croissante* [the argument from increasing automation], qui préconise une connexion éthiquement saillante entre l'autonomie (robotique) d'un agent artificiel et sa propension à nuire aux êtres humains. En ce sens, si l'autonomie d'un agent artificiel est suffisamment vaste ou sophistiquée, cela peut susciter le besoin d'une moralité artificielle au sein de ce dernier, et donc la création d'un agent moral artificiel.

Reste que ce statut d'agent éthique explicite demeure problématique, notamment puisqu'il ne nous aide aucunement à identifier *quels* aspects d'un agent moral humain doivent être simulés afin d'assurer un comportement idéal. Pire encore, il nous conduit même à un paradoxe : trop d'approximation des caractéristiques humaines, nous fait courir le risque de transformer les agents moraux artificiels en *patients moraux*, signifiant que certains comportements, voire même certains droits, leur seront naturellement dus de la part des agents humains. En revanche, pas assez de ces caractéristiques, et nous courons le risque de contrecarrer la possibilité même d'une moralité artificielle, et donc de construire des machines qui posent des risques éthiques aux êtres humains. Néanmoins, il semble que le point idéal entre ces deux extrêmes consiste à revendiquer l'approche du *maximaliste* : l'implémentation computationnelle d'une théorie morale particulière au plus haut degré de sa possibilité technique. Ce faisant, on évite non seulement la création de *personnes artificielles* qui perturbent le système moral humain, mais de plus, on pourra même créer des machines qui font preuve d'un comportement moral exemplaire, voire angélique [better angels of our nature], si du moins on accepte que les agents humains ne soient pas catégoriquement exemplaires dans leur prise de décision morale. Cela nous conduit enfin à une sorte d'ironie philosophique : que les entités qui ne sont pourtant pas des agents moraux, pourraient toutefois agir *mieux* que les agents moraux ; une thèse que nous allons analyser avec soin dans la deuxième partie.

Chapitre III : *hétéronomie, modularité et les agents moraux artificiels*

A travers les deux derniers chapitres, nous avons abordé diverses manières pour les agents moraux artificiels de différer des êtres humains. Le chapitre III vise à regrouper ses idées, en posant trois distinctions ontologiques. En premier lieu, et en suivant notre discussion de l'autonomie robotique, nous préconisons que les agents moraux artificiels existent catégoriquement dans un rapport d'*hétéronomie* avec un ou des agents humains. Cela implique que a) les machines n'ont pas d'autonomie télique [goal autonomy][11], et b) qu'un être humain (l'utilisateur principal, le programmeur, etc.) doit décider des fins que la machine poursuit à travers son action pratique. En ce sens, les robots deviennent des moyens exotiques pour les fins humaines.

Deuxièmement, et en suivant notre discussion de l'environnement d'un agent moral artificiel et de son *Umwelt*, nous préconisons que les agents moraux artificiels soient des agents *modulaires*, pratiquement et ontologiquement restreints par rapport aux agents universaux humains. Ces restrictions prennent trois formes différentes : des restrictions *extensionnelles*, qui limitent la gamme des buts pratiques que l'agent peut accomplir dans son environnement (i.e. une voiture autonome doit conduire des êtres humains, et non pas trier des colis amazon), des restrictions *agentives* qui limitent la gamme des entités que l'agent peut percevoir dans son environnement (i.e. une voiture autonome doit distinguer des piétons, des voitures ou des camions, mais elle ne doit pas reconnaitre la différence entre un combattant et un civil), et enfin des restrictions *intentionnelles*, qui visent à limiter les méthodes par lesquelles une machine peut répondre aux entités de son environnement en poursuivant ses buts pratiques. Cette dernière restriction traite la question des procédures de décision (éthiques) au sein des agents artificiels autonomes, et les restrictions intentionnelles sont donc équivalents à la moralité artificielle d'un agent.

[11] Dignum, 2019.

Enfin, nous introduisons un troisième type de distinction, relevant cette fois du type de relation qu'un agent moral artificiel peut entretenir avec les agents humains. Tandis que tout agent moral artificiel existe nécessairement dans un rapport d'hétéronomie avec les fins des agents humains, cette relation peut parfois influencer les détails de ses restrictions intentionnelles. En effet, il existe une large gamme de machines, que nous proposons d'appeler les *agents mandataires* [surrogate agents], qui agissent *au nom d'un individu spécifique* au sein de leur *Umwelt.* Cela revient à dire que ces agents, en faisant des courses pour un individu, en conduisant à sa place, ou bien en prenant des décisions variées pour le bien d'un individu, pourraient susciter des attentes en termes de *loyauté* ou de *partialité* de la part de cet individu, en vertu du fait qu'une portion de son autonomie décisionnelle soit déléguée à cette machine[12]. Cela nous mène à l'idée que l'intérêt personnel de l'utilisateur principal pourrait être vu comme un facteur moralement saillant de la moralité artificielle de certains types de machines. En revanche, il semble évident que tout agent moral artificiel ne tombe pas facilement dans cette catégorie d'assistant personnel, puisqu'une grande gamme de machines semblent accomplir des fins plutôt générales, ou du moins, des fins de personne en particulier. Nous proposons d'appeler les machines de ce genre les *agents distributifs,* puisque pour ces machines, l'intérêt personnel d'un individu quelconque ne semble pas avoir d'importance morale. Les robots opérant dans les contextes médicaux, ou au service des forces de l'ordre, semblent tomber dans cette deuxième catégorie.

Toutefois, force est de constater que les catégories d'agent mandataire et d'agent distributif n'admettent pas toujours de limites claires et strictes. En effet, il semble qu'un nombre important d'agents moraux artificiels, ne serait-ce qu'en vertu de leur nouveauté, tombent facilement dans les deux catégories à la fois, selon la perspective analytique que l'on adopte. Par exemple, la voiture autonome peut être perçue comme un agent mandataire selon son passager, mais pourrait atteindre le statut d'agent distributif pour les autres individus dans l'environnement routier, voire même pour la société en générale. Cela suggère que la recherche d'un contenu substantiel d'une moralité artificielle chez ce genre d'agents peut s'avérer difficile, puisque les attentes normatives des agents humains quant au comportement de ces agents ne sont ni uniformes, ni complémentaires.

[12] Johnson & Powers, 2008: Keeling et al., 2019: Millar, 2014.

Chapitre IV *: les contraintes techniques et la structure de la moralité artificielle*

Ayant terminé la première partie de notre analyse, nous nous penchons maintenant sur la structure et le contenu de la moralité artificielle. Dans la littérature, il existe trois grandes catégories—ou styles d'implémentation—d'une moralité artificielle : les approches dites 'top-down', 'bottom-up', et 'hybride'[13]. Les approches top-down, autrement appelés des 'systèmes experts', sont caractérisées par la mobilisation d'une théorie morale explicite, suivant un style de programme déterministe. Les approches bottom-up, à leur tour, mobilisent une gamme assez vaste d'approches stochastiques : l'agrégation des préférences morales d'une population suite aux études empiriques[14], l'agrégation et la classification des aspects moralement saillants d'un contexte décisionnel par les agents humains[15], ou encore l'apprentissage moral par renforcement. Les 'hybrides', quant à eux, représentent une catégorie assez vague, mais sont souvent caractérisées par la présence des aspects top-down et bottom-up, où les règles strictes d'une théorie explicite rendent plus transparents et plus prévisibles les détails décisionnels fournis par l'apprentissage bottom-up.

Il semble que nous retrouvons ici les mêmes distinctions et tendances que nous avons abordées lors de notre analyse ontologique. Cependant, il semble aussi que nous retrouvons les territoires respectifs du maximaliste et du minimaliste, si nous examinons ces approches non pas sous l'angle de leur style d'implémentation, mais plutôt selon leur *source de contenu moral.* Dans ce chapitre, nous nous penchons sur le cas du maximaliste et abordons le problème principal du minimaliste (le fait que l'acceptabilité morale des préférences sociétales soit douteuse) dans le chapitre V.

Le maximaliste, en mobilisant une théorie morale comme source de contenu moral, et surtout comme source de *procédure de décision éthique*, est confronté à trois problèmes distincts.

[13] Allen & Wallach, 2008: Allen, Varner & Zinser, 2000: Allen, Smit & Wallach, 2005.
[14] Bonnefon et al., 2016 : Awad et al., 2019.
[15] Conitzer et al., 2017.

Premièrement, le problème de ce que nous proposons d'appeler la *circonscription* d'une théorie morale [constituency of a moral theory]. Cela revient à dire qu'en vertu du fait que les agents moraux artificiels ne sont pas des agents moraux entiers (par leur manque de subjectivité, d'autonomie morale, etc.), il s'ensuit que tout effort d'implémentation d'une théorie morale demandera un degré important d'*interprétation* par le programmeur humain. Autrement dit, les théories morales s'appuient souvent sur des capacités ou des qualités mentales que les robots ne possèdent pas. Ce problème fait donc écho aux difficultés rencontrées dans le chapitre II, lorsque l'implémentation de certaines capacités métaphysiquement épaisses semblaient nécessaires pour le bon fonctionnement d'une théorie morale, mais semblaient cependant problématiques d'un point de vue éthique.

Deuxièmement, ce problème de circonscription entraine un problème encore plus profond, celui de ce que nous proposons d'appeler *la moralité sans pression* [pushless morality][16]. En deux mots : si l'on peut accepter que le fondement des échanges moraux entre agents humains soit caractérisé par un échange entre des agents moraux entiers, mais aussi des *patients moraux entiers*, le fait qu'un agent moral artificiel n'a pas de statut de patient moral perturbera fondamentalement la nature de cet échange, et par conséquent, la structure de toute théorie morale qui la présuppose. Le terme « pression » en ce sens revient à souligner comment la valeur d'un agent moral entier peut « mettre la pression » sur les exigences que la moralité peut lui imposer : il est en ce sens raisonnable par exemple qu'un agent moral entier ne soit pas dans l'*obligation morale* de dédier *toutes* ses ressources matérielles et temporelles au bien-être des plus mal lotis, puisque ses projets personnels, ainsi que son bien-être individuel, comptent parmi les aspects moralement saillants de sa prise de décision éthique[17]. Force est de constater qu'il n'en va pas de même pour les agents moraux artificiels, et que par conséquent, les exigences de la moralité ne connaissent vraisemblablement aucune limite chez eux. Autrement dit, les exigences morales venant du statut moral des êtres humains peuvent faire pression sur le comportement d'un agent moral artificiel, mais le statut d'agent moral artificiel ne peut ni réduire la sévérité de ces exigences, ni lui même faire pression sur les comportements des agents humains.

[16] Nozick, 1984, 401.
[17] Williams, 2011: Nagel, 2012.

Enfin, une certaine ignorance de ce problème de la « moralité sans pression » dans la littérature, a amené plusieurs auteurs vers une vision erronée de ce que nous proposons d'appeler la *place* de la moralité artificielle au sein du comportement général d'un agent moral artificiel. En effet, du côté de la robotique, il existe une supposition assez générale selon laquelle la moralité artificielle est un aspect *ponctuel* ou *épisodique* du comportement d'une machine[18]. Cela revient à dire qu'il est plausible que la poursuite de la totalité des buts pratiques d'un agent ne demande pas un encadrement moral, mais seulement la poursuite de ces derniers à travers certains contextes décisionnels. En effet, si l'on suppose que la moralité artificielle d'un agent est omniprésente dans son action pratique, on risque de nier non pas sa capacité de pression *morale*, mais plutôt de pression *pratique*, en générant des agents qui n'arrivent pas à accomplir leurs fins, faute de l'exigence sévère de la moralité. Pour fixer les idées, imaginons un robot caissier qui obéit strictement au seul principe de non-nuisance. Si un client veut acheter des cigarettes, il est probable que ce robot ne laissera pas passer cet achat : le tabagisme diminue la probabilité de survie de 0.0004%[19]. De plus, il est également probable que la plupart des objets en vente poseront problème au robot : les sacs de courses en plastique peuvent être nocifs pour les enfants, les briquets, et même pas mal d'aliments gras ou sucrés. Même si cet exemple est extrême, il illustre le fait qu'une application omniprésente d'une moralité artificielle peut contrecarrer le but même d'un agent moral artificiel.

Reste qu'il semble tout aussi difficile de définir la place idéale que devrait occuper la moralité artificielle au sein du comportement d'un robot. A cet égard, nous préconisons trois visions différentes de cette place : large, étroite, et modérée [wide, narrow, moderate]. Le cas d'une place large revient à celui d'une moralité omniprésente et n'est vraisemblablement utile que pour les agents moraux artificiels dont le rôle même est de prendre des décisions éthiques. Les robots médicaux et, peut-être, les robots de guerre semblent tomber dans cette catégorie. La vision étroite de la place de la moralité artificielle correspond aux cas où seuls les contextes décisionnels les plus dilemmatiques au sein d'un *Umwelt* particulier signalent le besoin d'une prise de décision éthique. Par exemple, les collisions inévitables d'un véhicule autonome relèvent de cette catégorie, où la prise de décision éthique n'est activée que lorsque toutes les actions possibles du véhicule mènent

[18] Dignum, 2019, 77.
[19] Leben, 2018. On imagine bien pourtant que ce chiffre n'est pas exact.

à une collision. Il est important de souligner que, dans le cas de cette vision étroite, l'agent moral artificiel doit être doté d'une capacité de *détection* de ce genre de contextes. Enfin, la vision modérée de la place de la moralité artificielle constitue une catégorie assez hétérogène, puisqu'elle ouvre la possibilité que *plusieurs formes* de moralité artificielle puissent être utiles au sein d'un même *Umwelt*. Prenons encore l'exemple des voitures autonomes. Il est relativement certain que les collisions inévitables constituent un bon motif pour l'implémentation d'une moralité artificielle, mais est-ce là toutefois toutes les situations pour lesquelles une telle capacité peut s'avérer utile ? Les conducteurs humains s'engagent souvent dans les situations de prise de risque à portée potentiellement éthique : la circulation autour des ronds-points, les interactions avec les piétons, voire même les changements de voie avec une visibilité sous-optimale. Ce que la vision modérée de la place de la moralité artificielle offre à la conception des voiture autonomes, c'est précisément cette opportunité d'appliquer des contraintes explicites et éthiques à ce genre de contextes. En ce sens, il serait possible d'implémenter une forme de moralité artificielle qui n'opère que dans les contextes de collision, et puis une autre forme, opérant sous des règles différentes, pour les situations de prise de risque. A travers l'exploration de ces trois problèmes, nous commençons donc à voir les possibilités exotiques qui se cachent au sein du concept d'une moralité artificielle.

Chapitre V : *l'acceptabilité et la moralité artificielle*

Tandis que la moralité humaine n'est pas communément influencée par des idées telles que l'adoptabilité de ses recommandations, ni la viabilité de son contenu vis-à-vis des préférences des partie prenantes industrielles, il semble qu'il n'en va pas de même pour la moralité artificielle. En effet, puisque la conception d'un agent moral artificiel passe par des étapes qui ne sont pas seulement théoriques et techniques, mais aussi juridiques et publiques, il semble que la conception du comportement moral de l'agent artificiel ne peut pas faire complète abstraction de ces facteurs. Cette idée demeure problématique pour le maximaliste, puisqu'il ne semble pas évident que des facteurs comme l'acceptabilité publique en elle-même portent une signification morale. Cependant, cette idée est beaucoup plus accessible au minimaliste, qui lui-même s'appuie précisément sur ce genre de facteurs dans l'élaboration d'une moralité artificielle. Néanmoins,

l'idée d'incorporer les contraintes d'acceptabilité au sein de la moralité artificielle est assez peu traitée dans la littérature, et constitue pour autant une tâche étonnamment difficile et subtile.

Dans ce chapitre, nous abordons trois facteurs différentes qui pourraient servir comme fondement pour l'acceptabilité d'une moralité artificielle : ce que nous appelons l'acceptabilité comme préférence morale, l'acceptabilité comme adoptabilité, et enfin, l'acceptabilité comme la viabilité institutionnelle. Prenons chacun de ces concepts à leur tour.

Le cas de l'acceptabilité comme préférence morale correspond précisément au programme du minimaliste : récolter des données sur les attentes normatives d'une population donnée, et s'efforcer d'en tirer une structure de moralité artificielle. En ce sens, le but de l'acceptabilité comme préférence morale est de découvrir la forme que prend la *moralité du sens commun* [common-sense morality] autour d'un *Umwelt* particulier. A cet égard, le projet de *Moral Machines* de MIT semble être un bon exemple de cette approche[20]. Cependant, cette approche rencontre des difficultés lorsque a) on cherche à établir un lien de ressemblance entre les résultats de cette approche et une théorie normative quelconque, ou b) lorsque l'on cherche à prendre une position normative par rapport aux résultats. Nous pouvons voir l'interaction de ces deux problèmes au sein du projet *Moral Machines* lui-même. En premier lieu, les résultats du projet semblent indiquer une sorte de 'dilemme social' entre deux préférences divergentes : les passagers qui semblent préférer des véhicules qui les protègent lors d'un accident, et une préférence plus générale pour la minimisation du nombre de blessés. Si l'on cherche à ramener ces préférences à des théories normales concrètes, nous voyons des points de similarité entre les préférences des passagers et une forme d'égoïsme ou de « déontologie[21] », alors que la préférence générale semble correspondre au principe utilitariste. De là, il semblerait non seulement que ces préférences sont foncièrement divergentes, mais de plus, que les préférences des passagers sont d'une qualité morale inférieure, en vertu du fait qu'une moralité artificielle programmée selon un principe égoïste accroitrait ostensiblement le nombre de blessés sur les routes.

[20] Bonnefon et al., 2016 : Awad et al., 2019 : Jaques, 2019.
[21] Frank et al., 2019: Shariff, Bonnefon & Rahwan, 2017 : Meder et al., 2019.

Le problème est que rien dans les données récoltées par le *Moral Machine Experiment* n'indique clairement cette opposition entre égoïsme et utilitarisme. Cela semble vrai en vertu du fait que la recommandation morale qui consiste à minimiser le nombre de blessés n'est pas exclusive de la pensée utilitariste, mais peut aussi bien être validée par des théories contractualistes, contractariennes, voire même déontologiques. De plus, certaines de ces théories (notamment le contractualisme) pourrait accorder une saillance morale à *plusieurs facteurs*, prenant par exemple le bien-être du passager comme une obligation spéciale qui contraint la maximisation de l'utilité générale. À la lumière de cette option, la position qui consiste à diviser cette forme de moralité de sens commun en deux camps nécessairement opposés, pour ensuite poser un jugement normatif sur la viabilité morale d'un des deux camps, semble erronée et superficielle. Il en résulte que les informations données par l'acceptabilité comme préférence morale, aussi précieuse soient-elles, sont néanmoins assez fragiles et limitées. Au mieux, nous pouvons n'espérer qu'une vision partielle des attentes normatives *locales* : au sein d'un *Umwelt* donné, et par le biais d'un type de contexte décisionnel spécifique.

Néanmoins, nous pourrions toutefois considérer que l'association entre données empiriques et théories normatives, aussi problématique soit-elle, n'efface pas entièrement les problèmes liés à l'acceptabilité générale d'une moralité artificielle. Cela semble particulièrement vrai pour le principe de minimisation du nombre de blessés : est-ce qu'il est vraiment socialement acceptable de protéger le passager, *même si* cela pourrait accroitre le nombre de blessés ? Ici, nous rentrons dans le territoire d'une deuxième forme d'acceptabilité : l*'adoptabilité* d'un agent moral artificiel. En deux mots, l'acceptabilité comme adoptabilité désigne les conditions nécessaires pour qu'un agent humain *accepte d'utiliser* un agent moral artificiel, plutôt que d'accomplir une tâche par ses propres moyens. A travers notre analyse, nous insistons sur le fait qu'il existe au moins deux seuils d'adoptabilité : un premier, relevant d'un principe de *parité comportementale*, qui souligne l'idée que le comportement d'une machine doit être qualitativement équivalent à celui de l'humain qui adopte la technologie, puis un deuxième, opérant par un principe d'*optimalité comportementale*, qui suggère qu'une machine doit agir *mieux* qu'un être humain dans son accomplissement de la tâche.

Cette discussion nous permet de fournir la réponse suivante à la question de la minimisation du taux de blessés : si l'on accepte qu'une certaine loyauté soit attendue par les passagers de la part des véhicules autonomes (selon le principe de parité ou d'optimalité comportementale), il semble qu'un véhicule programmé uniquement pour minimiser le nombre de blessés pourrait être *inadoptable*. Plutôt que de voir en cela une autre forme de dilemme social, nous pourrions faire une distinction entre le critère de justesse [criterion of rightness] générale d'un agent moral artificiel d'un côté, et la procédure de décision éthique d'une moralité artificielle. Autrement dit, si le but ultime des voitures autonomes consiste malgré tout en la minimisation du taux de blessés suite aux accidents de la route, il ne suit pas que chaque voiture *individuelle* doit être programmée de cette façon. En effet, il est probable qu'une moralité artificielle qui témoigne d'un certain degré de loyauté envers son passager soit beaucoup plus adoptable et pourrait néanmoins faire baisser le taux de blessés suite aux accidents de la route, du moins par rapport aux standards des conducteurs humains. De plus, les conditions d'adoptabilité sont très protéiformes : ce qu'il est nécessaire d'implémenter aujourd'hui pour assurer l'adoptabilité peut changer demain, en fonction de la publicité d'une technologie et une fois que la compréhension de cette technologie est arrivée à maturité.

Enfin, nous arrivons au troisième sens d'acceptabilité : la viabilité institutionnelle. En général, cette forme d'acceptabilité ressemble souvent à une adhésion aux principes liés aux problèmes éthique de la conception, tels que la transparence d'une machine ou sa redevabilité. Ici, nous choisissons cependant d'explorer un autre angle, celui du croisement entre d'un côté, des facteurs et des caractéristiques d'un environnement qui serviraient comme informations *utiles* pour une moralité artificielle quelconque, et de l'autre, la viabilité institutionnelle de la collecte et de la mobilisation de ces mêmes données. Nous explorons cette idée à travers l'examen d'un processus nécessaire pour toute approche maximaliste et minimaliste, celui que nous proposons d'appeler le *processus d'adduction morale artificielle* [the process of artificial moral uptake]. En effet, si l'on cherche à inclure un facteur moral telle que la vulnérabilité dans la moralité artificielle, il faut spécifier a priori quelles caractéristiques environnementales serviront comme *ancrage* pour ce facteur : l'âge d'un individu, ses vêtements, sa mobilité, son genre, etc. Très vite, on s'aperçoit que l'efficacité de la définition extensionnelle d'un principe ou d'un facteur s'obtient au détriment du respect de la vie privée des êtres humains ou de leur dignité. Cette confrontation entre l'éthique

décisionnelle et l'éthique de la conception atteint donc ici son apogée, dévoilant une sorte de problème de cadre d'acceptabilité [acceptability frame problem].

En effet, l'acceptabilité comme viabilité institutionnelle exerce une pression sur la précision de la prise de décision éthique, relevant de deux problèmes distincts. Premièrement, il existe ce que nous appelons le problème d'un *seuil d'ignorance morale* [threshold of moral blindness] qui occulte des informations qui, malgré leur utilité, semble d'être des connaissances inacceptables pour une machine. Nous songeons ici aux caractéristiques comme le niveau socio-économique d'un individu, ou son casier judiciaire. Ensuite, il existe aussi le problème de ce que nous proposons d'appeler les *appâts pervers* [perverse incentives][22], qui quant à eux, suggèrent l'idée que la publicité de certains aspects de la prise de décision éthique d'une machine peut engendrer des réactions nocives ou contreproductives chez les agents humains de son *Umwelt*. Par exemple, s'il est généralement connu que les voitures autonomes protègeront en priorité les cyclistes sans casque (au motif de leur vulnérabilité relative), il est possible que les cyclistes arrêtent de porter des casques, faute d'être ciblés par des voitures autonomes lors d'un accident[23].

Ensemble, ces deux problèmes occultent énormément d'informations utiles pour le bon fonctionnement d'une moralité artificielle. Il peut même arriver qu'une stricte adhésion à la viabilité institutionnelle empêche totalement une prise de décision éthique, comme cela semble être le cas avec la recommandation allemande pour les voitures autonomes[24]. Néanmoins, ces problèmes peuvent être atténués si l'on accepte de faire une distinction au sein du genre de faits qui sont pris en compte par ce processus d'adduction morale artificielle. Il est évident, par exemple, que certains faits ont un rôle *constitutif* au sein de la prise de décision éthique. La probabilité de survie d'un individu semblerait être constitutive d'une moralité sensible au bien-être des parties prenantes. De plus, il semblerait que certaines caractéristiques additionnelles pourraient affiner la détection du bien-être, telles que l'âge (général) et le genre, puisque ces éléments pourraient avoir une impact véritable sur la probabilité de survie lors d'un accident de voiture. A notre sens, il semblerait donc que des tels faits constitutifs relèvent des aspects objectifs du contexte décisionnel,

[22] Loh & Misselhorn, 2019.
[23] Goodall, 2014.
[24] Luetge, 2017.

et n'emportent donc pas de préférence ou de préjudice quelconque. De plus, ces caractéristiques sont typiquement détectables par la seule capacité perceptive de la machine elle-même, et en ce sens, ils ne constituent pas un empiètement sur la vie privée d'autrui, pas plus qu'un simple regard entre inconnus.

Toutefois, il existe toute une autre gamme de faits, que nous proposons d'appeler des faits *scalaires,* qui ne constituent pas des éléments objectifs d'un environnement, mais au contraire le genre de préférences externes que les problèmes éthiques de conception cherchent à neutraliser. C'est ainsi que la quasi-totalité des caractéristiques mobilisées dans le *Moral Machine Experiment,* afin de solliciter les préférences morales des individus, relèvent de cette catégorie de faits scalaires : l'occupation, l'âge, le casier judiciaire, etc. Si une moralité artificielle était basée sur ces faits scalaires, il semblerait que le mieux que l'on pourrait espérer d'une telle approche serait une forme de prioritarisme en faveur des chouchous de la société et non pas une *moralité* à proprement parler. De plus, la détection de ces faits ne passerait pas uniquement par le biais de la capacité perceptive de la machine, mais bien par une récolte de données éthiquement troublante qui poserait surement un problème substantiel pour le respect de la vie privée. A notre sens, une moralité artificielle basée uniquement sur des faits constitutifs pourraient plus facilement se glisser sous la voile d'ignorance imposé par la viabilité institutionnelle.

Chapitre VI : *la moralité artificielle et la théorie des valences éthiques*

À ce point, nous avons passé en revue les obstacles inhérents à la conception d'une moralité artificielle. Reste à identifier les défauts les plus importants chez les maximalistes et les minimalistes. Nous avons vu, par exemple, que l'approche minimaliste, aussi bien d'un point de vue computationnel que d'un point de vue de source du contenu moral, rencontre des difficultés sérieuses dans les problèmes éthiques de conception et d'acceptabilité. Nous avons aussi constaté que le maximaliste, quoique relativement sans reproche à la lumière de ces deux aspects, doit néanmoins reformuler toute théorie morale afin de l'implémenter dans une machine : il court alors

le risque de prôner une moralité que personne n'accepte et qui néglige le but pratique de cette machine même.

À la lumière de ces problèmes, il semble qu'un principe général régulant la bonne conception de la moralité artificielle soit nécessaire. Nous proposons pour ce faire ce que nous proposons d'appeler le *principe de l'irréprochabilité totale*. Sa justification est la suivante : si l'on peut accepter que l'intention originale de la mise en place des agents moraux artificiels, et donc la création d'une moralité artificielle, est de faire en sorte que les agents artificiels ne nuisent pas aux êtres humains dans les contextes de saillance éthique, il s'ensuit que la maximisation de la non-nuisance aux humains doit servir de règle générale de la moralité artificielle. Toutefois, la clef de cet argument tient dans la définition substantielle de « nuisance ». En vertu peut-être de la progression même de la technologie robotique, nous associons souvent l'idée de « nuisance » avec la notion de dommage physique, voire même du dommage létal chez l'humain. Cela ramène le territoire de la nuisance robotique au plus proche du territoire de la sécurité de l'utilisateur[25]. Est-ce là toutefois tous les sens pertinents du mot « nuisance » ? A notre sens, non. Comme nous l'avons vu, l'impact des décisions prises par les machines peut s'étendre bien au-delà du dommage physique : une machine peut décider qui aura un poste au sein d'une entreprise, qui sera libéré sur parole, et qui aura droit à une transplantation d'organe. Même si le dommage physique pourrait être un facteur moralement saillant dans la plupart de ces contextes, il est en revanche plus délicat de dire que ce sera *le seul* facteur qui ait une importance morale. Nous devons donc étendre notre concept de nuisance pour qu'il s'applique aussi à ces dommages non-physiques.

Pour ce faire, nous proposons de baser la moralité d'un agent artificiel sur les attentes normatives des individus quant à son comportement. Plus précisément, le principe de l'irréprochabilité totale nous invite à maximiser la responsivité d'un agent moral artificiel à ces attentes. Cela implique que l'agent doit s'efforcer de répondre de manière maximale aux valeurs dont les occupants humains de son *Umwelt* sont porteurs. En ce sens, un agent humain aura un grief légitime si, en décidant qui recevra un bénéfice quelconque, l'agent moral artificiel ignore la juste revendication de cet individu, optant plutôt pour la maximisation pure de l'utilité générale. Vu que ces revendications seront naturellement basées sur la forme de moralité du sens commun,

[25] Van Wynsberghe & Robbins, 2019.

le principe de l'irréprochabilité totale revient à demander à ce que la moralité artificielle épouse ce sens commun, plutôt que d'en intégrer seulement quelques aspects. En ce sens, si une société voit le rapport entre le passager et une voiture autonome comme moralement saillant, la moralité artificielle des voitures autonomes doit en faire autant.

Qu'est-ce que cela implique pour le maximaliste et le minimaliste ? Il semblerait que le maximaliste verra sa gamme de théories morales acceptables diminuer encore, puisque le principe de l'irréprochabilité totale semble interdire toute théorie qui *révise* la moralité du sens commun. Pour le minimaliste, en revanche, il semble que la forme de moralité du sens commun fournie par les études d'acceptabilité ne peut passer ce cap d'irréprochabilité que s'il peut éviter les problèmes de seuil d'ignorance morale et les faits scalaires. De plus, le minimaliste doit trouver un moyen d'organiser ces préférences morales au sein d'une procédure de décision cohérente, sans pour autant tomber dans une forme de prioritarisme basée sur les faits scalaires. À notre sens, il est possible d'accomplir cela, en associant les points forts du maximalisme et du minimalisme.

Cette idée nous conduit à la théorie des valences éthiques. En effet, plutôt que de préconiser *une* procédure de décision éthique basée soit sur des préférences morales sociétales, soit sur une théorie morale quelconque, la théorie des valences éthiques propose un système de *mitigation des revendications* [claims] particulières à chaque individu dans l'environnement de l'agent moral artificiel. Premièrement, cela implique que le fondement de ces revendications pourrait changer en passant d'un *Umwelt* à un autre, et deuxièmement, cela implique que la façon même de mitiger ces revendications peut changer : selon les changements d'acceptabilité publique d'une technologie, et selon les besoins particuliers des parties prenantes variées.

Conceptuellement, la théorie des valences éthiques est composée de trois éléments distincts : les revendications [claims], les valences, et les profils moraux. Les revendications, pour commencer, suivent les permutations d'un facteur moralement saillant d'un *Umwelt* à travers les agents humains qui s'y trouvent, et servent comme des injonctions morales *pro tanto*. Par exemple, si le bien-être est le facteur moralement saillant opératoire, chaque revendication individuelle sera plus au moins forte en fonction des faits constitutifs comme la probabilité de survie de chacun. Il en résulte que les individus n'auront pas la même puissance revendicative aux yeux de l'agent

moral artificiel. Le but de la théorie des valences éthiques est cependant de répondre maximalement à toutes ses revendications, et si impossible, de faire en sorte que l'agent moral artificiel donne la priorité à la revendication la plus forte de son environnement dans sa sélection de l'action.

Cette responsivité aux revendications de chacun est cependant troublée par un deuxième élément conceptuel : les valences. Le but des valences est de capturer l'acceptabilité inhérente de chaque revendication dans un *Umwelt*. Il est donc ici question d'incorporer l'acceptabilité comme préférence morale, puisque le fondement d'une valence individuelle relève des données empiriques de ce genre. En ce sens, la valence d'un individu peut être plus ou moins forte, moyennant la façon dont ses particularités individuelles répondent aux critères d'acceptabilité sélectionnés. L'enjeu, dès lors, est de construire une étude d'acceptabilité qui n'intègre que les faits scalaires les moins inoffensifs, ceux qui risquent de passer sous le voile d'ignorance de la viabilité institutionnelle. Par exemple, si dans le cas des véhicules autonomes, les valences étaient basées non pas sur une préférence morale générale, mais plutôt sur une caractéristique relationnelle de l'*Umwelt* comme la *vulnérabilité,* il est probable que les faits scalaires qui en résultent seraient presque de l'ordre des faits constitutifs : on pourrait retrouver des catégories de personnages comme les cyclistes, les camionneurs, les piétons ou les enfants. De là, il est possible d'organiser ces catégories dans une sorte d'hiérarchie, où les plus vulnérables selon l'acceptabilité publique (les piétons, ou les enfants) auront la valence la plus forte. En ce sens, l'action de guider la recherche de l'acceptabilité vers des facteurs de valeur éthique plausible, et non pas *vers* la prejudice, pourrait atténuer au moins une portion des inquiétudes liées à l'utilisation des faits scalaires.

De plus, le rôle que jouent ces valences est loin d'être superficiel. Pour fixer les idées, imaginons un cas où un véhicule autonome doit choisir entre le sacrifice d'un adulte ou d'un enfant lors d'un accident inévitable. Il s'avère que les revendications des deux individus sont équivalentes, si les probabilités de survie de chacun sont par hypothèse les mêmes. Sans les valences, la voiture serait dans l'obligation de choisir aveuglément entre ces deux individus, puisqu'il n'est ni possible de satisfaire les deux revendications, ni possible de privilégier la plus forte des deux. En revanche, avec les valences, la voiture pourrait opter pour le sacrifice de l'adulte,

puisque la valence de l'enfant (du moins si elle est basée sur la vulnérabilité) serait surement plus forte que celle de l'adulte. Il semblerait donc non seulement que ce choix est moralement préférable à l'alternative d'un choix aveugle, mais de plus, que ce choix soit guidé par les attentes normatives de la société, le rendant de ce fait plus acceptable.

Cependant, il n'est pas non plus vrai que les valences auront catégoriquement le dernier mot sur la prise de décision éthique. La théorie des valences éthiques préconise de privilégier la *revendication* la plus forte en cas de conflit. En ce sens, si les revendications de l'enfant et de l'adulte étaient différentes, la voiture aurait choisi de sacrifier celui qui revendiquait le moins dans son environnement. De plus, si les valences sont configurées selon des critères assez vagues comme le type d'usager de la route, il est possible que l'acceptabilité des revendications individuelles varie assez peu au sein d'un contexte décisionnel particulier. Néanmoins, l'atout principal d'une valence relève de sa capacité à faire pencher la balance vers les choix ostensiblement acceptables dès lors que les revendications de chacun sont les mêmes.

Enfin, la mitigation de ces revendications et de ces valences est elle-même réglée par le profil moral opératoire, où chaque profil fournit une procédure de décision éthique distincte. En ce sens, il est possible d'accommoder une grande variété de procédures de décision, y compris celles des théories morales classiques comme l'utilitarisme ou le principe rawlsien de maximin. Néanmoins, ces procédures ne font usage que des revendications des individus et elles risquent donc d'être moins acceptables que d'autres configurations possibles. A la place, nous pourrions songer, dans un esprit un peu plus « contractualiste », à des sortes de *compromis* possibles entre les revendications qui se trouvent le plus souvent en conflit. Dans le cas du véhicule autonome, cela reviendrait à faire une séparation catégoriale entre les revendications des passagers d'un côté, et les revendications de ceux qui se trouvent à l'extérieur du véhicule, de l'autre. Cela nous permet de concevoir des profils moraux à *seuil,* où le passager pourrait accepter un certain degré de dommage physique pour sauver la vie d'un piéton, mais pas au point où il risque d'en être gravement blessé. Dans cet esprit, il est dès lors possible d'incorporer l'acceptabilité comme adoptabilité au sein de la prise de décision éthique d'un agent moral artificiel.

La théorie des valences éthiques échappe ainsi à un grand nombre de problèmes soulevés au fil de nos analyses. Premièrement, elle évite les problèmes de circonscription morale, puisque la théorie elle-même est conçue non pas pour un être humain, mais bien pour un agent moral artificiel qui doit répondre maximalement aux pressions morales de son environnement. Il en va de même pour le problème d'interprétation d'une théorie morale, puisque l'agent moral artificiel, à travers la théorie, est doté d'une *perception morale écologique* des revendications de son environnement et n'a pas besoin de raisonner pour sentir sa responsivité à ces éléments. Deuxièmement, les trois visions possibles de la place de la moralité artificielle peuvent être réconciliées au sein de la théorie des valences éthiques, soit par le biais d'un même profil moral, soit à travers plusieurs profils selon le type de contexte décisionnel. Troisièmement, enfin, il est évident que notre théorie considère l'acceptabilité publique comme un facteur substantiel de la prise de décision éthique, mais elle évite de se baser entièrement sur ces préférences. En ce sens, les exigences de la moralité occupent toujours une place centrale dans la moralité artificielle, mais cette place est modérée par les attentes normatives des individus.

Chapitre VII : *la théorie des valences éthiques et les voitures autonomes*

La théorie des valences éthiques a été élaborée au sein du projet AVEthics, un projet pluridisciplinaire regroupant des philosophes, des roboticiens et une psychologue autour du défi de la conception d'une moralité artificielle acceptable pour les véhicules autonomes. De plus, un des buts principaux du projet étaient de fournir une approche computationnelle complète, qui serait prête à être implémentée dans les véhicules autonomes des partenaires industriels de l'institut Vedecom. La théorie, et surtout son application computationnelle appartiennent donc au projet entier, et aux membres qui le composent. En ce sens, le chapitre VII présente l'esquisse d'une implémentation computationnelle de la théorie des valences éthiques accomplie par les membres roboticiens, implémentation dont le détail serait trop long à aborder dans ce résumé. Néanmoins, le chapitre VII sert aussi à faire apparaître les enjeux principaux des véhicules autonomes et ce sont eux que nous allons présenter brièvement ici.

Premièrement, il est notoire que le véhicule autonome, par rapport à la plupart des agents moraux artificiels, bénéficie d'une publicité inédite. En effet, il existe déjà des prototypes sur les routes américaines, ou des fonctionnalités comme « autopilot » dans la plupart des voitures de luxe actuelles, et leur arrivée générale sur le marché commercial est anticipée pour 2030. Quant aux considérations éthiques des véhicules autonomes, elles impliquent deux choses : premièrement, l'arrivée de ces véhicules relève d'un processus d'automation *incrémental*. Cela veut dire que les voitures futures seront automatisées pas à pas, plutôt que d'arriver d'emblée avec une automation complète. Néanmoins, le but de ce processus est d'arriver à ce que l'on appelle communément le niveau « 5 » d'automation, qui implique que les voitures peuvent conduire dans toutes sortes de circonstances, sans jamais déléguer la tâche de conduire à l'humain. Deuxièmement, l'arrivée générale des voitures autonomes relève aussi d'une *temporalité* incrémentale. Cela implique que les voitures autonomes, pendant longtemps encore, doivent interagir avec des conducteurs humains. Ce n'est que dans un lointain avenir que la majorité des véhicules sur la route seront complètement autonomes. En ce sens, la réactivité et la capacité sociale d'un véhicule autonome doivent s'adapter à cet agenda temporel, surtout lorsqu'il s'agit de la communication avec les piétons, mais aussi du point de vue des attentes normatives autour de la prise de décision éthique de ces véhicules. Il est donc très important de préciser où l'on se place sur ses échelles incrémentales lorsqu'on évalue l'impact éthique de ces machines. A cette fin, notre analyse se concentre sur le début de ce processus incrémental, à un moment où les voitures autonomes de niveau 5 doivent toujours interagir avec les conducteurs humains.

Deuxièmement, il est important de souligner que la vertu principale de l'arrivée anticipée des voitures autonomes, pour la quasi-totalité de ses parties prenantes, consiste en une baisse considérable des morts liées aux accidents routiers, une baisse qui pourrait toucher les 90%. Les justifications de cette attente sont multiples : principalement, on estime qu'une voiture autonome, n'étant ni distraite, ni sous l'emprise de l'alcool, et n'ayant pas non plus une tendance à la conduite agressive, pourrait facilement passer le cap du principe de parité, voire même d'optimalité comportementale. Sinon, d'autres motifs plus technologiques sont présents : les voitures autonomes pourraient communiquer entre elles afin d'éviter des accidents, ou bien les voitures autonomes en vertu de leur capacité perceptive impressionnante, pourraient voir les accidents à temps pour les éviter. Néanmoins, il serait erroné de voir les voitures autonomes comme des

conducteurs parfaits. La raison en est qu'en vertu du mélange entre les conducteurs humains et robotiques, les accidents risquent de persister du moins pour le futur proche. De plus, le comportement très craintif et robotique de ces voitures, ainsi que leur éventuelle panne mécanique, fera en sorte que des accidents, même rares, continueront à se produire.

C'est peut-être pour cette raison que les voitures autonomes ont suscité autant d'intérêt philosophique. En effet, si une voiture autonome est fatalement confrontée à des collisions inévitables, et que de plus, elle est suffisamment rapide pour prendre une décision délibérée plutôt que de réagir instinctivement dans ces contextes, il semblerait dès lors que la voiture puisse être considérée comme une instance réelle du célèbre *dilemme du trolley*, conçu par la philosophe Philippa Foot et plus tard, par Judith Thomson. Il n'est pas donc étonnant que la plupart de la littérature traite ce problème du trolley dans le cadre des véhicules autonomes. Cependant, il est important de séparer, d'un côté, l'utilité de ce problème du trolley comme une expérience de pensée abstraite pour l'élaboration d'une moralité artificielle, et de l'autre, ce problème du trolley comme représentation exacte des conditions matérielles des accidents de voiture autonome. Il semble évident qu'une expérience de pensée philosophique ne reproduit pas les conditions d'un accident de voiture au pied de la lettre, mais il s'agit d'une autre question tout entière que de demander si ces différences porte une saillance morale. À notre sens, le dilemme du trolley offre un cadre adéquat pour révéler des attentes normatives ou pour détecter les facteurs et caractéristiques moralement saillants de la prise de décision d'un véhicule autonome.

Par conséquent, il semble important de traiter le sujet de la moralité artificielle elle-même dans le contexte des véhicules autonomes et du genre d'approches qui ont été proposées. Premièrement, il est notoire qu'il existe un consensus dans la littérature autour de l'importance morale de la probabilité de survie dans la prise de décision éthique. En effet, la quasi-totalité des auteurs mobilise ce facteur dans leurs approches. Cependant, il existe peu de facteurs autre que celui-ci qui semblent avoir une réelle importance dans la littérature : seules une obligation spéciale pour le bien-être du passager et une distinction entre les agents humains impliqués et non impliqués dans un accident ont reçu une attention notable. Enfin, la prise de décision éthique semble aussi susciter de l'intérêt chez les maximalistes. En effet, à part l'approche des valences éthiques, la totalité des approches relèvent des théories (morales) préexistantes : l'utilitarisme, la déontologie

kantienne, le principe de Maximin, la doctrine de la nécessité légale, ou bien des approches contractariennes. Ceci demeure étonnant, puisque les études empiriques très vastes comme le *Moral Machine Experiment* démontrent qu'il y a un doute que l'acceptabilité de telles approches.

Enfin, le sujet émergent principal de l'éthique des voitures autonomes semble être celui de la responsabilité (légale ou morale) de la prise de décision de ces véhicules. D'un côté, ce problème touche le sujet plus général du respect du code de la route. Il peut sembler évident qu'une voiture doit respecter la loi, mais le problème est que ce respect même, s'il est trop rigide, porte des risques d'accidents et de mécompréhension des conducteurs humains. D'un autre côté, il semble compliqué de programmer une voiture pour qu'elle ne respecte pas la loi de façon délibérée et intentionnelle. De plus, la justification d'une telle approche semble délicate, puisqu'il serait tout à fait possible d'augmenter la qualité légale de la conduite humaine, plutôt que de baisser celle des voitures autonomes. Enfin, il semble plausible que la responsabilité morale elle-même porte une saillance morale dans la prise de décision éthique, et les approches émergentes de moralité artificielle s'efforcent souvent de prendre cette facteur en compte.

Conclusion

Le monde actuel est de plus en plus numérisé. Nous vivons des vies entières dans des espaces virtuels différents, et dépendons de plus en plus des prédictions, des décisions et de l'aide des agents artificiels. À l'époque où nous avons entamé la recherche qui a conduit à cette thèse, les problèmes éthiques liés à l'automation croissante étaient surtout d'ordre spéculatif. Peu de chercheurs professionnels prenaient ce problème au sérieux, et peu d'industriels trouvaient que l'éthique des technologies avait une place au sein de la stratégie commerciale d'une entreprise. Aujourd'hui, les « principes éthiques de l'IA » sont au bout de la langue du monde entier, face aux problèmes que cette technologie a posé pour la liberté individuelle, le consentement, l'autonomie, et la démocratie. À l'heure actuelle, il semble que le sujet de l'éthique des agents artificiels s'est échappé de sa tour d'ivoire théorique, et suscite l'attention des politiciens, industriels, législateurs, voire même des sociétés entières.

En ce sens, s'il reste une place pour la pensée philosophique au sein de ces problèmes éthiques, il ne s'agit sûrement ni d'une place spéculative, ni d'une place alarmiste. Les

philosophes, lorsqu'ils se penchent sur la question des agents artificiels, doivent se rendre compte de l'impact même de cette activité dans le monde pratique et réel. En ce sens, l'idée d'implémenter une moralité kantienne au sein d'un robot personnel, aussi intéressante soit-elle sur un plan théorique, aura un vrai impact sur la vie de la personne qui interagit avec ce robot. Il semble donc pertinent de se demander non seulement ce qu'il est *juste* d'ajouter au comportement de ces agents, mais aussi ce qui est désirable et acceptable d'y ajouter. Du point de vue de l'univers, il pourrait être vrai que les robots agissant comme des agents utilitaristes parfaits pourraient rendre le monde meilleur, mais il n'est pas évident que ce monde meilleur plaira aux utilisateurs, voire même aux constructeurs de ces technologies.

C'est dans cet état d'esprit que la théorie des valences éthiques a été conçue. Sa volonté de marier les aspects dits « minimalistes » et « maximalistes » dans la moralité artificielle d'un agent moral artificiel n'est pas le résultat d'une passion syncrétiste de nature purement spéculative. Mais elle s'inscrit, du moins l'espérons-nous, dans un vaste courant de recherches publiques et pluridisciplinaires autour de l'acceptabilité du comportement robotique, une recherche qui se poursuivre bien au-delà de ce travail doctoral.

Introduction

Consider the following account of a day in the life of a typical human, Robert:

Robert starts his day the way most of the modern developed world does, by checking his phone. There, he notices a reminder he's left for Mrs. Pedersen's birthday, the mother of his best friend. He promptly decides to call his friend, and leaves a message wishing them a lovely day together before heading off to work.

Once at work, Robert settles into today's task: selecting potential candidates for a new position that has opened up at his office. Robert considers a number of applicants, all of whom have diverse merits and backgrounds. After comparing the different applicants against the prescribed hiring profile, he decides to call three of them back for an interview, all of whom are male. Over lunch, he shares his decision with his co-workers, and one of them promptly files a complaint with human resources for gender discrimination.

Upon leaving the office, a different co-worker asks Robert if he can catch a ride home. Robert agrees. Once on the road, the pair quickly encounter a parked delivery truck on a two-lane street. Robert is eager to get home, and knows the delivery could take some time. So, he decides to maneuver around the truck, and as he does so, a pedestrian darts out in front of the car. In a split-second decision, Robert swerves to avoid the pedestrian, careening into a streetlamp which totals his car, and leaves his co-worker with a bad case of whiplash.

Finally, after dropping his co-worker off at the hospital, Robert walks back to his apartment, picking up his mail on the way. There, buried amongst bills he hasn't paid and ads he doesn't want, he finds an invitation to Mrs. Pedersen's funeral, who died the previous Sunday. Robert instantly regrets the cheery phone call he made to his best friend.

By all accounts, Robert has had a particularly bad day: he was privy to incomplete information, applied faulty administrative procedures, and was the victim of almost comical bad luck. For many of us, Robert's day would inspire a bit of empathy, or a heart-felt pat on the back. In any case, very few of us would call Robert's day *unethical*, or his actions *immoral*.

But what would change if our hero was not Robert, but *Robot?* How would our moral appraisal shift if an autonomous vehicle decided to give its passenger whiplash, or if a decision-assistance program selected only male candidates to hire? How much would we appreciate automated birthday notifications from dead relatives on our Facebook feeds? Chances are that our tolerance for *Robot*'s actions would be comparatively low, even if *Robot* and Robert are privy to the same information, and are subject to the same events beyond their control. Patently, while *Robot* and Robert are both agents in the same world, the type, quality and scope of their actions are different, as are the types of moral attitudes we can hold in their regard.

The aim of this thesis is to mitigate the difference between Robert and *Robot*, or between human agent and *artificial agent* when they act in society. Specifically, this thesis will attempt to identify the ways by which *Robot*'s actions could be rendered *acceptable* to people like Robert; both in terms of the kinds of actions *Robot* can and cannot perform, and the principles, values, and procedures which could form the basis of *Robot's* decisions.

As we will quickly discover, there are many elements that set artificial and human agents apart. While their situations may be similar, they are nevertheless not equipped with the same cognitive tools and processes, nor do they share the same breadth of knowledge about themselves, the world, and their place within it. Likewise, their capacity for self-explanation, justification and accountability are wildly dissimilar, almost as divergent as their capacity to be responsive to the claims, rights, preferences and intentions of the individuals they interact with. Artificial and human agents often don't even look alike, as the former can take myriad forms: from cars and drones in physical space, to algorithms and chatbots in cyberspace. Yet, while they are heterogenous and myriad, the robots of the world must still abide by the same social, legal and moral constraints as Robert when they act in society.

Thankfully, we are not alone in this pursuit. Indeed, there exists an ever-swelling rank of roboticists, psychologists, legal experts and philosophers which have risen to this challenge, giving way to the emergent field known as *machine ethics*. These thinkers have taken up the creed of devising systems of rules and procedures which will govern the development of artificial agents along desirable ethical lines; lines which extend into the reasoning processes of the robots themselves. To this end, even if the object of machine ethicists can vary extensively across a variety of application cases, they are united in a common cause: to ensure that emergent and increasingly autonomous artificial agents do not harm the human agents they engage with, or the society in which they act. Characteristically, this concern for the moral risks of automation has led to the development of a particular kind of machine: artificial *moral* agents.

These agents are peculiar among the larger pool of intelligent machines for two reasons: first, they are destined to act in contexts where moral agency appears to be required, or what we will call *ethically salient contexts*, and second, they are equipped with a particular type of programming to respond to the moral aspects of their environment, what we will call *artificial morality*. By these lights, a robot exploring the Moon's surface is likely not an artificial moral agent, but an autonomous drone exploring the surface of a war zone likely *is*. Taken this way, the ethical objections we might level against machines hardly appear novel, since the ethical ramifications of emergent technology have long since been noticed and nuanced by a host of

thinkers. In this sense, nuclear weapons, industrial machines, and even television or the internet have all been the subjects of technophobic appeals to the moral risks they pose for society. However, the heart of the concern that drives the development of artificial *moral* agents is not a question of ethical *impact*, but rather a question of *decisional capacity*.

In effect, artificial moral agents will need to make decisions which themselves have ethical impact. An autonomous drone may decide who to target or execute, a decision-assistance system might decide who to hire or fire, and an autonomous vehicle, in an unavoidable collision, must decide how to crash. In each case, it is difficult to trace the answers to these questions back to a pure question of function optimization—since coding the best way to kill enemy combatants seems fundamentally dissimilar from coding the best way to clean a floor. In this sense, machine ethicists have seen fit to incorporate various *ethical constraints* into the decisions of these machines, building artificial moralities that promote some ends over others, and forbid some actions in pursuit of others; thus rendering these agents responsive to aspects such as the value of human life or personal autonomy, and thereby curbing their harmful impact on human social spheres.

Machine ethicists, however, have encountered one major problem in this pursuit: it is not entirely clear to *which* ethical standards machines such as these ought to be built. Worse still, this problem is two-sided. On one side, the ontological differences between humans and machines threaten what is perhaps the most obvious form of artificial morality: the simulation of human moral reasoning processes in artificial moral agents. In this sense, the hardware and software of machines serves as a very shaky foundation for robust and complex metaphysical concepts such as moral responsibility, blame and praise, or moral motivation, subjectivity and consciousness. Indeed it seems that this ontological difference bars artificial moral agents from the very activity of moral agency. In this sense, any attempt to replicate the human moral mind in these agents will be superficial, syntactic and incomplete. On the other side, even if these ontological hang-ups can be overcome, there remains the even more daunting question of the *type* of moral mind we ought to implement. Robots do not have their own essential moral character, and it follows from this that we must somehow decide which principles, axioms, values or ends they ought to pursue. Given that we are neither subject to a global consensus concerning the principles of the ideal society, nor

are we able to ascertain universal moral truth, a vexing amount of arbitrariness plagues the design of artificial morality.

Faced with this challenge, the response of the machine ethics community has typically been that of implementing computational models of traditional ethical paradigms into the programming of artificial moral agents. In more philosophical corners, these proposals have been either *experimental*—to see if a moral theory is truly implementable—or normative and relatively *universal*—that all machines ought to abide by these standards. Then, if these authors get their way, some or all machines will engage with society as perfect maximizing consequentialists, dutiful Kantian agents, empathetic Smithian machines, or virtuous Aristotelian partners. Another, emergent course of action consists in imparting a moral education on artificial moral agents, by training them to recognize patterns in societal moral preferences, or teaching them how to behave by moral example. Then, these crypto-Rousseauian machines will behave like perfect cultural relativists in their environments, modifying their moral minds in function of the wisdom of the crowd.

Of course, neither approach definitively solves the challenge of artificial morality. The first type of response which we might call *maximalist*, while surely maximizing the potential for moral behavior, nevertheless only imparts snippets of the full picture of human morality onto machines. Then, when programmed this way, machines might act in a supererogatory and admirable manner in response to some moral aspects, but be completely blind and coldly indifferent to others. The second type of response, which we could call *minimalist*, likely generates more well-rounded artificial moral agents, but the moral quality of its actions will always be capped by the ethical quality of the data from which it learns. In this way, the maximalist can accuse the minimalist of a dangerous naturalistic fallacy, transforming the sometimes ethically dubious descriptive behavior of humans into ostensibly laudable artificial moralities. In response, the minimalist can accuse the maximalist of imposing a rather arbitrary vision of the Good Life on the design of artificial morality, exhibiting a type of moral responsiveness that no human agent truly asked for or desired.

In this thesis, we will attempt to show that this binary choice between maximalist and minimalist approaches is unfounded, since both operate on the assumption of what we might call the *mimetic fallacy*: that the simulation of *human* moral behavior, principles and processes is what is required for acceptable moral responsiveness in *artificial* moral agents. Instead, we will defend a more *exotic* view, one that attempts to blend the best of both of these options, and chases after an acceptability-oriented ideal. We will claim that the original impetus of artificial morality—to prevent harming humans across the behavior of machines—is not best matched by a principle of moral perfection, but rather a principle of total *irreproachability*: that artificial morality ought to render a machine maximally responsive to the moral expectations of the human agents with which it interacts. In this sense, while the scope of this principle may be universal, its application most certainly is not. Rather than opt for a universal approach to the design of artificial morality, we will attempt to carve out an approach with a more *modular* appeal, one which looks to the moral expectations specific to a given human social sphere, a given task environment, and an explicit and bounded purpose for which an artificial moral agent is implemented.

To shoulder all of this, however, we must trace the evolution of artificial moral agents from the very beginning of their story. Indeed, the first part of this thesis is dedicated to a thorough appraisal of the shift from artificial agent, to artificial moral agent, and investigates the necessary dissonance between human and machine ontology. In chapter I, we focus on two themes: agent and environment. Our main argumentative line will be that of exploring how fluctuations in the characteristics of machines (i.e., whether they are embodied, whether they act in dynamic environments) impacts the way in which they are programmed, and the type of moral impact they will likely have. This will cause us to search out a workable definition of the baseline concept of an *artificial agent*, as well as that of an agent program, a decision context, and the type of world knowledge these aspects require. Additionally, our discussion of hardware and software brings us to an orthogonal, but nevertheless very important ethical consideration in the design of intelligent artefacts: that of what we will call *ethical design concerns*. These concerns do not directly relate to the morality of a machine's decisions in the way that artificial morality does, but rather, act as ethical constraints on the *design* of even an amoral machine. These concerns typically take the form of design principles such as accountability, transparency, and responsibility, and have in recent years become *very* hard soft norms in technological development indeed. In this sense, while

ethical design concerns might not direct the types of principles or values that a machine ought to expound through its decisions, they do afford a robust second-order appraisal of the ethical impact certain design choices in artificial morality will have.

In chapter II, we will investigate the principal fulcrum upon which the dissonance between human and artificial agents rests: the contentious and ambiguous concept of *machine autonomy*. In the machine ethics literature, there is perhaps no other subject—save perhaps for the dubious concept of machine *intelligence*—which has caused so much frustration and misunderstanding, and it is here that we will come to bear with the interdisciplinarity inherent to the design of artificial morality. The concept of autonomy in machine ethics has often admitted of *levels*, ranging from minimal agents who are subject to ethical concern only in terms of their design and placement in human social contexts (landmines will serve as our archetype here), to highly sophisticated, superintelligent agents, whose epistemic superiority may even warrant a type of moral deference on the part of human agents. This autonomy scale can then be seen to track multiple factors: the decisional division of labor between humans and machines, the increasing computational complexity of machines, the increasing approximation of the human moral mind in machines, and their slow encroachment into the territory of human moral agency and the status of personhood. Here then, we will address how the will to avoid certain *ethical* ramifications in highly complex machines—such as the attribution of rights or moral responsibility for actions—impacts the design of artificial moral agents. We will also come to grips with a tangential corner of machine ethics, one which asks *whether* artificial moral agents could or should become true moral agents in the human sense of the term. This will lead us to two rather paradoxical ideas: first, that artificial moral agents must fail to satisfy the conditions of the standard view of moral agency if their design is to be ethical, and second, that this in no way precludes the possibility that machines might be *more* ethical than humans, resembling what we will call the *better angels of our nature.*

In chapter III, we provide our own ontological taxonomy of artificial moral agents, riding the coat tails of our discussion of machine autonomy. In effect we will argue that current artificial moral agents—or what we will call level 4 explicit ethical agents—are subject to a host of necessary characteristics. First, we will claim that these machines exist in a necessary relationship of *heteronomy* with a human programmer or user, and their purpose-oriented ontology is to provide

an exotic, relatively autonomous means of achieving this person's ends. Second, we will claim that artificial moral agents are *modular*, admitting of three restrictions: *extensional*, pertaining to the limited ends these machines can pursue, *agentive*, pertaining to the types of objects and subjects these machines can recognize and interact with, and *intentional*, pertaining to the types of policies and principles they can apply in the achievement of their goals. Finally, we will posit a further and loose distinction between what we will call *surrogate* and *distributive* forms of artificial moral agent. Surrogate agents, as the term implies, act on behalf of, or in the interest of, some human agent; typically a principal user. This opens up the possibility that in achieving this user's ends, the machine is in some meaningful way morally constrained by the first or second-order interest of this user. Distributive agents, on the other hand, are beholden to no human agent in particular, and are instead constrained by more general moral principles and policies.

In part II of the thesis, we depart from our discussion of ontology and move concretely into the territory of artificial morality. Here, our challenge will be to investigate, and eventually answer, what we will call the *diamond question of machine ethics*. This question sits at the intersection between acceptability, engineering and moral philosophy, and asks how we can resolve the computational and philosophical limitations of artificial morality in such a way as to provide acceptable behavior in a modular artificial moral agent. In so doing, we will spend much time with the two rival characters of artificial morality design: the maximalist, and the minimalist.

In chapter IV, we focus on the technical limitations of artificial morality. We begin by exploring what are likely the main approaches to artificial morality in the machine ethics literature: top-down, bottom-up, and hybrid implementations. These concepts, too, admit of different interpretations across the engineering and philosophical corners of machine ethics. In brief, the top-down approach to artificial morality is the natural territory of the maximalist, since it is only through this route that explicit normative theories can be implemented into a machine. The bottom-up model then easily lends itself to the minimalist, since the aggregation of the wisdom of the crowd is easily achieved in this model. Through our analysis, we will start to pick apart the moral high ground of the maximalist, by pointing to two main problems. First, what we will call the problem of *constituency*, which relates to the idea that given a machine's necessary failure to satisfy the conditions of the standard view of moral agency, any moral theory we seek to implement

must admit a relatively high degree of *interpretation*. Most damningly, the fact that artificial moral agents have no subjectivity, leads us to the idea that they have no moral *push*, meaning that they are not owed any specific moral behavior on the part of human agents. This is problematic for moral theory, in so far as it characteristically presupposes an interactive ebb and flow—or push and pull—between moral agents of equal moral status. Thus, for a moral theory to be implementable, it must be rendered *pushless*. Second, we will come to terms with what we will call the *place* of artificial morality. In effect, moral theories not only tend to presuppose moral agents as their subjects, but also tend to suppose that these agents are of a *universal* variety, meaning that they mobilize their moral nose across a wide variety of decision contexts. In machine ethics, this has led to a presupposition of universalism that is not only unwarranted, but that also can lead to a dubious and 'exceptionless' type of moral behavior in artificial moral agents, one which tends to neglect the practical purpose for which these agents were built, and which humans expect them to achieve. The place of artificial morality then denotes the *scope* of a machine's application of its artificial moral nose.

Chapter V moves away from these rather theoretical limitations, and into the concrete and confusing territory of what we will call *acceptability* constraints. These take the form of limitations or recommendations that an artificial morality ought likely to satisfy if it is to yield an acceptable machine worth purchasing or approving for public use. There are three kinds: first, acceptability as moral preference seeks to map the moral expectations of the users (or society) inherent to a given machine's environment. The point of this investigation is to discover what we will call the *shape of local common-sense morality*, or the amorphous set of action-guiding recommendations, or morally relevant features and factors that people may expect their machine to be responsive to. Secondly, there is the concept of *acceptability as adoptability*, which specifies the conditions under which a human agent may elect to use a machine, or delegate some of his decisional autonomy to it. Finally, the conflict between artificial morality and ethical design concerns is directly addressed by what we will call *acceptability as institutional viability*, a problem we will explore through what we will call the *process of artificial moral uptake*. This process consists in a designer (or a machine's) detection and classification of environmental features which could be seen to carry moral importance. Unfortunately, many such ostensibly useful and admissible features are obscured by a *threshold of moral blindness*, which occurs when an ethical design concern 'blocks'

the detection of a feature such as the age, gender, or socio-economic status of an individual. In this sense, we end chapter V with a sketch of an *acceptability frame problem,* which any account of artificial morality ought likely to solve.

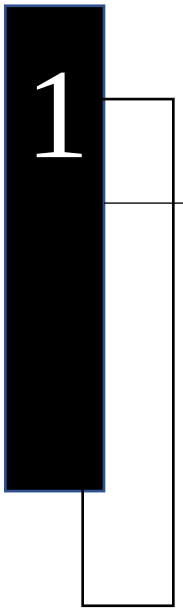
Finally, in chapter VI, we take this acceptability frame problem, and the problem of pushless morality, and attach it to our overriding design objective of satisfying the principle of total irreproachability. We will hold that the maximalist suffers from a problem of legitimacy, in so far as his ardent will to perfectly expound an arbitrary moral theory comes at the cost of disappointing many human agents, and potentially thwarting the threshold of moral blindness. The minimalist, however, is at a loss to organize his vast web of collected preferences in a way which guarantees moral responsiveness, and which doesn't fall into the morally dubious trap of simply prioritizing the empirically proven darlings of his sample population. In light of these shortcomings, we will propose our own solution, the Ethical Valence Theory, which paints artificial morality as an exercise in claim mitigation, where the goal of the agent is to maximally respond to all of the claims of its environment, and when this is not possible, to respond to the strongest claim. The theory itself revolves around three separate conceptual elements: claims, valences, and moral profiles. Claims in this theory attempt to track what normative ethics appears to require in machine behavior, based on the constitutive facts of the context (e.g., that an autonomous vehicle threatens human welfare, and thus individual claims track these fluctuations in welfare). Valences, on the other hand, are seen to reflect the more *scalar* facts which flow from acceptability studies, and the minimalist's ideal. Since each individual in the machine's decision context is afforded both a claim and a valence of varying strength, both normative and descriptive ethics have a role to play in the machine's responsiveness. Finally, moral profiles decide the decision procedure of the machine, or how these claims and valences should be weighed, prioritized or responded to. This yields a highly flexible approach to artificial morality, which is able to accommodate a wide variety of moral views, societal expectations, and stakeholder demands.

Finally, in chapter VII, we apply the Ethical Valence Theory to a specific type of artificial moral agent, autonomous vehicles. In effect, the theory itself was developed within an interdisciplinary research project, the AVEthics project, dedicated to the elaboration of acceptable

ethics policies in autonomous vehicles. This implies not only that the Ethical Valence Theory was developed with both public and industrial support, but also that it was elaborated with the intention of being fully implementable in fully autonomous vehicles. Thus, in this chapter, we first address some of the more general aspects of autonomous vehicles, and then move on to a computational implementation of the theory.

Taken together then, this thesis should be viewed as a pragmatic attempt to devise a form of artificial morality that honors not only what morality requires of intelligent machines, but what human beings ostensibly expect from them. The through line of this thesis then consists in this tension between, on one hand, the theoretical and philosophical wisdom which is required to understand what morality requires, and on the other, a more practical humanism borne of collaboration with various academic disciplines, and of the will to ensure that machines interact with humans in familiar, respectful and comfortable ways. In some sense, the prospect of designing 'perfect' artificial moral agents has led many to dream of a world in which machine decisions can improve or even correct the moral character of society, a thought which seems natural given the carefree atrocities that human beings have taken to committing. But it is questionable whether such a bizarre form of technological solutionism has a place in the moral corners of our lives, and more importantly, whether we yet know what that solution is.

CHAPTER



Artificial Agent & Environment

Today's society is home to two types of agents: human and *artificial*. Together, they contribute to the flourishing of society through *action*, by selecting goals or ends, and deliberating about the means by which to accomplish them. In the modern world, these actions can be highly complex and collaborative, involving the agency of multiple human and artificial agents (AA). Take for example a human agent, Martha, who wishes to become a world-class painter. Her goal (to paint well) requires a number of means, all of which are provided by a combination of human and artificial agency. She might for instance require a set of brushes and some paint, both of which are designed by human agents, and likely constructed by artificial agency via various industrial machines. Then, she might purchase these materials from an online marketplace such as Amazon.

There, a vast array of brush sets and oil paints are available for purchase by the human agents that are selling these products, and her choice between them may, in turn, be influenced by the recommendations of Amazon's algorithms (a type of artificial agency) or the reviews of other human users. The product she selects is then packaged and sent to her doorstep through a combination of human agency—the warehouse supervisors, coordinators, and workers, and eventually the mailman—and artificial agency: the vast logistical algorithms of the warehouse, the warehouse robots, and the digital logistics system of the post office. Finally, she is able to hone her craft, track her progress and develop an audience through the use of content hosting websites like YouTube, auto-didactic learning websites such as Masterclass, and social media platforms like Facebook and Instagram; all of which rely on an impressive degree of collaborative agency between human and artificial agents. Thus, if Martha wishes to become a world-class painter in 2020, it is virtually impossible for her to achieve her goal without the dovetailing cooperation of many hundreds of human and artificial agents.

This extensive apparatus on which Martha depends to achieve her goal is known as a *sociotechnical system*[1], denoting an environment in which human and artificial agency combine to further human ends, or provide goods and services[2]. While the sociotechnical system in which Martha acts is relatively novel, in so far as she must rely on many budding types of technology such as online marketplaces and social media platforms, sociotechnical systems themselves are hardly a phenomenon reserved to the 21st century. Since the dawn of the industrial age at least, humans and machines have collaborated to efficiently further human goals: from Ford's automobile assembly line and nuclear power plants, to hospitals, stock exchanges and the world wide web. It would seem, further, that the appeal of sociotechnical systems, from the very beginning, was always one of efficiency through delegation: through the handing-off of various tasks, duties or roles to technological systems, human agents were able to better address the goals they set out to achieve, be they individual or collective.

[1] Dignum, 2019: Winfield, 2015.
[2] To this end, (Pitt, 2011) describes technology itself as 'humanity at work', wherein all technological artefacts receive input from users and transform this into an output, on behalf of humans, and in pursuit of human interest.

The nature of this delegation, and the types of ends it affords to human agents, likewise constitute complex phenomena of long-standing philosophical interest. To this end, many thinkers of the 20th century were already able to see the transformative power of technology in society: positively contributing to human flourishing by "…spawning new ends (worthy or frivolous) from the mere invention of means…" or still yet, "…establish[ing] *itself* as the transcendent end."[3] Similarly, the notion of delegation and deference to machines itself was subject to critical analysis, since even if it provided efficient avenues for human flourishing, it may, in its wake, generate an over-dependence on technology[4], increasing human frailty and fallibility, while decreasing self-sufficiency, human connection, or self-awareness[5]. Thus if the intuitive appeal of technology was in some respects a question of pareto-optimality—that, ceteris paribus, the efficiency gained through technological artefacts and systems was better for some and worse for none—it nevertheless came at the price of some degree of *ethical risk.* At minimum, with delegation and efficiency came feelings of impotence and redundancy, and maximally, that fast-paced technological progress would paint basic human experience and communication as something of an anachronism.

Despite these concerns, the 21st century has opened its arms wide to the promise of machine delegation and deference. 'Better than human' technologies[6] have cropped up in a myriad of human social spheres, from abstract tasks like calculation, prediction and data-analysis, to concrete applications in law, healthcare, the military[7] or the traffic environment[8]. Importantly, the 'automation' of increasingly numerous aspects of the human social and political sphere has been accompanied in recent years by ever greater levels of ethical risk and failure[9]: the threat of 'push-button wars' and robotic arms races[10], the damage done to global democracy through misinformation campaigns and new-wave data analytics companies such as Cambridge Analytica; even the slew of 'driverless accidents' causing many a fender-bender, and occasionally human

[3] Jonas, 1979, 38.
[4] Bradshaw et al., 2013: Dignum, 2019.
[5] Turkle, 2017.
[6] Abney, 2012
[7] Arkin, 2009.
[8] Lin, Bekey, & Abney 2008, 69.
[9] Moor, 2005.
[10] Spiekermann, 2015: Cummings, 2006, 35: Sparrow, 2007

fatalities on America's roads[11]. Defenders of the benefits of these technologies—who often have a hand to play in their production and commercialization—have lamented the media's targeting of these ethically dubious 'outlier' cases, focusing instead on the comparatively greater benefits these technologies are purported to offer to society[12]. Familiarly, the cogency of their arguments revolve around claims concerning the dangers and sub-optimalities of the human performance of these tasks: humans are frail, emotional, biased, unreliable, slow-moving or generally *dangerous* to other humans whenever they perform certain roles in society[13], and thus technological delegation and deference, while not without its own risks, is still an improvement over the status quo.

But beyond these purported benefits and frightening outlier cases lies a deeper truth about the pace and sprawl of modern automation. While it is something of an eternal truth that machines are designed to assist, collaborate and improve upon the human attainment of ends, the social environments in which they are implemented nowadays seem to require an *ethically sensitive* form of collaboration. If autonomous vehicles are deployed on public roads, they may encounter unavoidable collision scenarios where human lives may be sacrificed. If autonomous weapons are deployed in battle fields, they may need to make ethically salient decisions balancing loss of life and the achievement of mission aims. A robotic healthcare assistant may need to decide how to distribute medication when not every patient can be adequately treated, and even a chatbot may enter into ethically dubious territory when it parrots the biased, discriminatory and insensitive language of the users with which it interacts. Unlike the form of ethical risk which preoccupied the philosophers of the previous century, this new form does not immediately relate to the generation of novel human ends, or the principle of complexity and 'automation surprises'[14]. It is

[11] The most paradigmatic example is perhaps the death of Elaine Herzberg on March 18th, 2018, when an Uber test vehicle (a Volvo XC90) struck Ms. Herzberg as she was crossing Mill avenue in Tucson, Arizona, outside of the designated pedestrian crosswalk (Stern, 2018: Griggs, 2018).

[12] Tesla, for instance, has maintained that their cars are 'safer' than human drivers, citing one accident every 2.87 million miles driven, whereas the NTSTA's data shows that an accident between human drivers occurs every 436,000 miles. (Davies, 2019).

[13] An exhaustive example of this line of reasoning can be found in Ronald Arkin's account of the pitfalls of the human soldier. He shows how many undesirable human characteristics (including, but not limited to emotion, tunnel vision, the fog of war, and the 'lack of combative spirit') can be 'corrected' out of a military offensive with the use of autonomous technology (Arkin, 2009, 29-48).

[14] Bradshaw et al., 2013. 'Automation surprises' result from the explosion of features, options and modes created by the implementation of technology, generating new types of demands, errors, and paths towards failure. In a similar vein, the principle of complexity, according to Bradshaw et al., captures the idea that machines, humans and macro cognitive work systems are fallible, and errors are therefore systemic. New

not an ethically loaded critique *of* a sociotechnical system. Instead, it is a truth of the *location* and *contextual reality* of a sociotechnical system; it is because humans wish to build sociotechnical systems, of great complexity, in the spheres of law, war, transportation and healthcare that the nature of the tasks, roles and ends of these fields themselves have ethical import, whether or not they are performed by a human or a machine.

Perhaps this is superficially true of more historical examples of sociotechnical systems. It is evident, for instance, that technology has for a long time played a significant role in the military sphere, and some of its applications—such as nuclear arms development—have clear ethical salience, regardless of the degree of human-robot collaboration that brought them about. But what is quite authentically novel about modern sociotechnical systems has more to do with the nature of the tasks that are delegated to machines, and correspondingly, the degree of oversight that human agents retain. If the environments within which humans intend to deploy sophisticated technological artefacts require some degree of ethical responsiveness, they often require that this responsiveness be visible in the *decisions* that these machines make. Having chosen to delegate the task of driving to a machine for instance, as is the case with autonomous vehicles, it is then the *machine's* task to 'decide how to crash', in a real-time traffic environment with human lives on the line. Similar parallels can be drawn in healthcare and the military, where task-delegation to machines has reached the decisional level in contexts of ethical importance. Thus, while it is evidently necessary to design the parameters and structure of modern sociotechnical systems in an ethically sensitive way, some social environments require the design of an additional *decisional capacity* within the technology which likewise honors ethical constraints. Put simply, in some application contexts, it is not enough for machines to be ethical simply in virtue of the spaces and roles in which humans have placed them, but further, a certain on-board *decisional* ethics is required in these machines if they are to correctly perform their tasks[15].

problems are associated with human-machine coordination, which can lead to breakdown, or the obscuring of the information necessary for human decision-making.

[15] In many respects, this claim mirrors a number of salient distinctions in the field of machine ethics, such as that of an implicit ethical agent, versus an *explicit* ethical agent (Moor, 2006), or that of ethics *in* design and ethics *by* design (Dignum, 2019). We will take up these distinctions at significant length farther on in this chapter, and in the next.

This nuance is not lost on the vast majority of the experts and researchers who consider these questions, often hailing from the field of what has alternatively been called 'roboethics'[16], computational ethics[17], science and technology studies[18], or most commonly, *machine ethics*[19]. While it is certainly fair to maintain that these fields aim to assess the general ethical impact of technology on society, the primary focus of these authors nevertheless tends to revolve around a particular kind of machine. In the broadest sense, these authors are concerned with what have been called *artificial agents*[20] (AA), denoting "…a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future"[21]. This rudimentary definition clearly captures the decisional aspects that make artificial agents a worthy subject of ethical concern and interest: namely, the 'sensing' and 'acting on' an environment in pursuit of 'its own' agenda. However, while artificial agents generally may pique the interest of machine ethicists, not every artificial agent, defined in this way, corresponds to the characterization of ethically salient sociotechnical systems that we have given above. An artificial agent in a meat-processing plant, for instance, may be able to act on and sense its own environment (by detecting and subsequently rendering animals), and it may pursue its own agenda—separating veal, pork and beef into appropriate areas of the factory. Yet, while this artificial agent is inevitably a part of a sociotechnical system, and while its activities understood in a large sense may carry ethical import for some (say, animal rights activists, vegans, or the labor force the artificial agent replaced), its decisions *themselves* are not obviously ethically salient. For this to be true, we would need to adjust the nature of the sociotechnical system: by, say, specifying that this meat-processing plant was the last to operate on a strictly carnivorous island, where not all islanders were able to purchase the meat they desired, and in addition to animal rendering and sorting, the robot was given the task of deciding which islanders received which meats. Here, the decisional aspects of the artificial agent's role carry ethical salience: it will

[16] Wallach & Allen, 2008.
[17] Floridi & Sanders, 2004.
[18] Johnson & Miller, 2008.
[19] Dignum, 2019: Allen, Smit, Wallach, 2005: Sullins, 2009: 2006.
[20] We have elected to use the term 'artificial agent' throughout the course of our analysis in keeping with the most common terminology from the philosophical tradition, see for example (Sullins, 2009) and (Franklin & Graesser, 1996). In most circumstances what is meant by 'artificial agent' is equivalent to a range of other terms: intelligent systems, autonomous systems, intelligent machines, intelligent agents, intelligent artefacts, etc.
[21] Franklin & Graesser, 1996, 25.

need to decide, independent of direct human control and supervision, which islanders receive meat, and which are left to starve.

In the literature, artificial agents that are placed in such ethically charged contexts, and who are *intentionally* designed to make such ethically-loaded decisions pertain to a special category: artificial *moral* agents (AMAs), "…the class of entities that can be involved in moral situations, for they can be conceived as…moral agents (not necessarily exhibiting free will, mental states or responsibility, but as entities that can perform actions…for good or evil)…"[22]. In comparison to simple artificial agents who may not act in situations of ethical salience then, AMAs present an additional design challenge for engineers: the machines themselves will need to be rendered ethically responsive, through explicit programming, planning and analysis by engineers. The rationale behind this supplementary design need is an intuitive one: "If multipurpose machines are to be trusted, operating untethered from their designers or owners and programmed to respond flexibly in real or virtual world environments, there must be confidence that their behavior satisfies appropriate norms. This goes beyond traditional product safety…[they] must also be 'cognizant' of possible harmful consequences of [their] actions, and [they] must select [their] actions in light of this 'knowledge', even if such terms are only metaphorically applied to machines".[23]

This 'step beyond product safety' and into moral territory goes by many names in the literature: machine morality[24], moral algorithms[25], ethical decision procedures[26], ethics policies[27], ethics settings[28], and perhaps more generally, artificial morality. In this chapter, our focus will be on the analytical subtleties of the transition from artificial agent to artificial moral agent, and accordingly, we will not yet venture into the details of the different engineering procedures and responses to this 'moral territory'. To this end, we will simply use *artificial morality* as a blanket term denoting a specific decisional portion of the design of an artificial agent, one that deals expressly with moral responsiveness to an ethically salient context, such as those we have cited

[22] Siciliano & Khatib, 2008, 12.
[23] Wallach & Allen, 2008, 17.
[24] Wallach & Allen, 2008.
[25] Leben, 2018.
[26] Keeling et al., 2019.
[27] Keeling, 2017: 2019.
[28] Millar et al., 2017: Lin, 2014a.

above. Accordingly, an artificial *moral* agent *necessarily* possesses some form of artificial morality, while many simply 'artificial' agents do not. Put another way, we might hold that the transition from AA to AMA passes through two conditions, one internal and one external. Externally, the environment of implementation of a specific artificial agent must carry some ethical salience, which we can now more formally describe as an environment wherein *moral value* is a part of the structure, function, or reality of the sociotechnical system, regardless of whether a human or a machine is made to respond to it[29]. Internally then, for an AA to be an AMA, it must be equipped with a way to identify, process, mitigate or generally *consider* this moral value in its decision-making, and the method by which it achieves this we will call *artificial morality.* Thus, for an artificial agent to become an artificial moral agent, it must both (1) operate in an environment which carries ethical salience, and (2) be equipped with a decisional procedure which allows the agent to consider human value in its decision-making, what we have called *artificial morality*.

With these somewhat loose bearings, the chapter will take the following form: we will first attempt to discern an operational and accurate definition of an artificial agent, and in so doing, reflect on the relationship between agent and environment. Then, in the second section, we will take a look at the concept of an agent's environment from an engineering-oriented perspective, and further hone our definition of an artificial agent. Finally, in section III, we will explore, rather topically, the overarching normative landscape within which all artificial agents are designed, and attempt to discern how this environment of ethical design comes to affect the types of design decisions an engineer can make.

[29] We have left the substantive elements of 'moral value' intentionally vague at this point in our analysis. Indeed, what can count as an ethically salient feature of an environment is hardly a closed question, and it is one that we will consider seriously and at length in the second portion of this thesis. If some precision is needed, common elements of 'moral value' in socio-technical systems may be elements like risk of human harm, risk of the thwarting of human interest, or certain duties and responsibilities that are owed to different individuals in virtue of their role, behavior, or relationship to the robot or other human agents; the operative assumption being that adequate robotic behavior in these context requires responsiveness to precisely these contextual features.

1. *Defining Artificial Agents*

Given our previous analysis, the term 'artificial agent' may already conjure images of humanoid-type robots, highly advanced technological artefacts that may handle jobs as complex as meat triage and resource management on a strictly carnivorous island. To be sure, this is one example of an artificial agent, but one can easily come across many other typifications in modern life. Indeed, artificial agents are operating behind the scenes of an impressive array of life's activities: from the chatbots that prowl the twitterverse or the home assistants that order groceries on command, but also 'smart appliances' that monitor food consumption and quality, or decision-assistance systems that provide reasoning support for legal arguments and precedent. The most evident characteristic that brings all these various entities together is their status as *technological artefacts*: "…purpose built artefacts designed, commissioned, and operated by human beings"[30]. These machines of various complexities and purposes are all the products of human intentions[31], non-natural entities built to further some human end, be it practical (to monitor grocery consumption) or experimental (to understand the nature of human intelligence or social cooperation)[32].

The disparity (both ontological and practical) between these various artificial agents, however, has much to do with the complicated job of correctly defining the limiting cases of what can count as an artificial agent. Some authors, for instance, take the concept of an artificial agent very broadly, maintaining that entities such as computer programs and computer viruses, thermostats and landmines can all count as some (minimal) type of artificial agent[33]. From the late 1980s until about 2010, this minimal characterization was popular. Intuitively, much of this popularity is likely owed to the absence of many of the more complex machines we see today, and

[30] Bryson & Kime, 2011, 1. More emphatically: "there is in fact no question about whether we own robots. We design, manufacture, own and operate robots. They are entirely in our responsibility. We determine their goals and behavior, either directly or indirectly through specifying their intelligence, or even more directly by specifying how they acquire their own intelligence. But at the end of every direction lies the fact that there would be no robots on this planet if it weren't for the deliberate human decisions to create them." (Bryson, 2010, 3).

[31] Johnson & Verdicchio, 2017: Johnson & Miller, 2008.

[32] Among the more complex artificial agents that we will address in the latter parts of this chapter, this distinction between practical and experimental aims will become important.

[33] Sullins, 2006: Franklin & Graesser, 1996: Johnson & Miller, 2008: Floridi & Sanders, 2001: 2004: Brustoloni, 1991.

perhaps more deeply, the need to account for the ethical changes that were taking place throughout the information and communication technologies (ICT) boom of that time period[34]. We can see an example of this broad interpretation of an artificial agent in the seminal textbook 'Artificial Intelligence: A Modern Approach': "…An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators"[35]. Quite clearly, rudimentary entities such as landmines and thermostats both meet these loose criteria, since a landmine 'senses' pressure, and a thermostat 'senses' a drop in ambient temperature. Furthermore, both of these basic entities 'act' on this information, and in a crude way 'decide' to explode[36], or engage a heating or cooling system.

The problem however, was that loose definitions such as these failed to capture useful distinctions between technological artefacts, which over the years, became increasingly salient to philosophical and scientific analysis[37]. Two of these characteristics were *autonomy* and *intelligence*, where intelligence was often construed as a capacity for reasoning, or context responsiveness, and autonomy, a capacity for independent, purpose-oriented action. Furthermore, these two capacities, within the context of artificial agents at least, often developed together, as reciprocal capacities that jointly ensured efficient action in artificial agents. Taken in roughly chronological order, we can see the definitional dovetailing of these two characteristics as they apply to the case of artificial agents:

> **Definition 1**: "Autonomous agents are systems capable of autonomous, purposeful action in the real world"[38].

[34] Indeed, some of the earliest academic conferences on machine ethics (such as the 'Ethics Comp' of the 1990s) aimed in large part to determine whether the supposed ethical problems of emerging technology were *novel* in any meaningful sense, or whether they were not simply a new configuration of familiar moral tensions and trade-offs. The seminal 'uniqueness debate' in machine ethics, or what Herman Tavani calls the 'computer ethics is unique thesis', revolves precisely around this question (Tavani, 2002). See also: Johnson (1985/ 2001): Maner, 1996.
[35] Russel & Norvig, 2013, 34.
[36] Asaro, 2006.
[37] Franklin & Graesser, who provide perhaps the most thorough taxonomy of artificial agents (from which some of our definitions are borrowed), further attribute this vagueness to the inherent complications of making absolute characterizations of real-world concepts: "The only concepts that yield sharp edged categories are mathematical concepts, and they succeed only because they are content free. Agents 'live' in the real world (or some world), and real-world concepts yield fuzzy categories" (1996, 24).
[38] Brustoloni, 1991.

> **Definition 2**: "Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed"[39].
>
> **Definition 3**: "Intelligent agents are software entities that carry out some set of operations on behalf of a user or another program with some degree of independence or autonomy, and in so doing, employ some knowledge or representation of the user's goals or desires"[40].
>
> **Definition 4**: an artificial agent is "…a hardware or (more usually) software-based computer system that enjoys the following properties:
> — autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;
> —social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language;
> —reactivity: agents perceive their environment…and respond in a timely fashion to changes that occur in it;
> —pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative"[41].
>
> **Definition 5:** "Intelligent agents are those that are capable of flexible action in order to meet their design objectives, where flexibility includes the following properties: (1) Reactivity: the ability to perceive their environment, respond to changes that occur in it, and possibly to learn how best to adopt those changes. (2) Proactiveness: the ability to take the initiative in order to fulfil their own goals. (3) Sociability: the ability to interact with other agents or humans"[42].

Starting with the notion of autonomy, the progression of the concept moves through an interesting set of stages. In definitions (1) and (2), 'autonomous' appears simply to denote an absence of human control. (1) appears to take autonomy as *purposeful* independent action, while (2) clearly places this purposeful and independent action within the context of achieving a set of goals set out by the *designer*. In definition (2) therefore, it is clear that an artificial agent's goal directed behavior is not chosen by the AA itself. Neither definition (1) nor (2) makes explicit mention of intelligence, beyond what may be intuitively required for 'purposeful action'. (3) in

[39] Maes, 1995, 108.
[40] Franklin & Graesser, 1996, 23.
[41] Wooldridge & Jennings, 1994, 2.
[42] Dignum, 2019, 10.

turn, shifts the choice of goal to the user, where the agent is acting on *behalf* of the user in the pursuit of *his* goals or desires. This likewise requires a certain internal interpretation of these goals and desires on the part of the AA, which paints the need for "some" degree of intelligence as a necessary condition of an artificial agent's independent and purposeful action[43]. Clearly however, the type of intelligence required in (3) is of a representational kind[44]. In definition (4) however, both autonomy and intelligence appear to take on a more socially oriented character, where purposeful and independent action appears to require the capacity for social interaction and proactivity. Quizzically, neither the role of the designer nor user is explicitly mentioned in (4), leaving the question of who chooses the aims of the artificial agent somewhat unanswered. To this end, despite the near 25-year gap between the publication of these definitions, (5) appears to echo (4) both in terms of its emphasis on social intelligence and interactivity, and on its vagueness concerning who sets the agent's goals. In (5), there is no mention of a user, but further, there seems to be some discrepancy between an agent's 'meeting [its] design objectives', and its 'ability to take the initiative in order to fulfil [its] own goals'. Are these two claims mutually incompatible?

Superficially, if we understand 'design objectives' to be nothing more than the goals the designer imagines for the artificial agent, then these claims are incompatible. Sensibly, there is only *one* agent doing the goal-choosing, either human or artificial. But we can avoid this definitional dead-end by drawing attention to a third characteristic which features prominently in the definitions above: the concept of *environment*. In effect, four of the five definitions make use of the term 'environment' in substantial ways: (1) specifies that an artificial agent acts in the 'real world' (a type of environment), (2) further clarifies that this environment is complex and dynamic, (4) posits that an artificial agent must 'perceive' its environment, responding in a timely fashion to changes which occur within it, and that beyond responsiveness, the agent must also 'take the initiative' *within* its environment to achieve its goals. Finally, (5) echoes responsiveness and

[43] Definition (3) is a classic example of what we will come to call the 'surrogate' model of artificial agency. It will be introduced in chapter 3.

[44] A good example of this type of representational intelligence and its relationship to independent action is the 'Belief, Desires, and Intentions' model of computer programming, a classical approach in artificial intelligence (Dignum, 2019, 18). Beliefs, for the artificial agent, pertain to its knowledge of the world (classically, in first-order representational logic), its desires pertain to its system goals, or the attainment of some end state, and finally, its intentions pertain to the potential plans the agent has to attain this end state or goal.

perception of the agent's environment, and further specifies that the agent may 'learn how best to adopt' the changes that occur within it. The concept of environment, therefore, is central to our understanding of artificial agents, in so far as the perception of, the acting within, and the responsiveness to an environment is the central pathway through which an agent's capacity for autonomy and intelligence can manifest itself. In other words, an artificial agent is not simply autonomous in the abstract, but rather autonomous within a given environment. Similarly, an artificial agent is not simply intelligent, but acts intelligently within its environment: by perceiving it, responding to it, acting (or interacting) within it, and learning from it.

The notion of environment, then, beyond being central to the definition of artificial agents, can help erase the apparent contradiction of the agent's 'design objectives' and the agent's ability to 'take the initiative in order to fulfil [its] own goals'. The simplest way to accomplish this is by looking at the *scope* of an artificial agent's environment. In all of the practical examples of artificial agents we have mentioned thus far—including the meat processing robot, the chatbot in the twitterverse, autonomous vehicles or autonomous weapons—*none* of the artificial agents can be seen to operate across an infinite set of contexts. An autonomous vehicle does not engage with users on twitter, and an autonomous drone does not decide which islanders deserve which meats. Instead, a specific artificial agent is designed to act in a *specific context* or *set of contexts*, which together provide the outer limits, or *scope* of the environment within which it is able to act and interact, to which it responds, or from which it learns[45]. Importantly, the artificial agent does not accidentally stumble into its specific environment. Instead, it is *intentionally* placed there by a human agent, and is intentionally designed to function *in* that restricted environment by a human agent. In the literature, this feature of artificial agents is captured by the concept of *purpose-oriented ontology:* their internal and external features are designed so as to perform a number of tasks in a pre-determined environment across time[46], a concept which flows quite naturally from the status of artificial agents as technological artefacts.

[45] The environmental specialization of artificial agents is not only relevant to the ontological investigation of 'what can count as an AA', it also has great import for types of moral behavior the machine can perform, and the types of behavior which may be expected of it. We will touch on these moral aspects of environmental specialization in the next chapter, and what we will come to call the 'modularity' of artificial moral agents will feature prominently in our discussion of artificial morality in part II.

[46] Franklin & Graesser, 1996. The authors further explain that situatedness in a specific context is in fact a condition of artificial agency itself, since an agent who is removed from its pre-determined environment

Philosophically, we can frame the concept pairing of a purpose-oriented ontology and a pre-determined environment through the concept of an *Umwelt,* a term coined by the biologist Jacob Johan von Uexküll to denote the world of unitary experience in which non-linguistic animals live[47], a 'soap bubble[48]' which traces the outer limits of animal life, providing a restricted practical realm for which a given species is tailor-suited. In detail, von Uexküll's concept of *Umwelt* relies on a specific perspective of the essence of animal nature: "…the animal is a vital subject whose core activities are perceptual and operational. Hence, an animal *Umwelt* is everything that the animal can perceive and do. *Umwelt* is the synthesis between what the animal perceives in its environment…and what the animal can do about what it perceives"[49]. The link between perception and action, in turn, is provided by what von Uexküll calls 'marks of significance' (*Merkmalträger),* objects in the animal's environment that prompt specific actions. The animal is able to differentiate between objects that bear marks of significance and those that do not, and in this way, its agency is *restricted* by the significant features of its environment.

More famously perhaps, Martin Heidegger, in his lectures on the fundamental concepts of metaphysics, repurposed the concept of *Umwelt* to denote the mode of being of animals, as entities who are 'poor-in-the-world' (*weltarm),* and surrounded by an intrinsic 'disinhibiting ring' which prescribes what can affect or occasion their behavior[50]. This condition is in stark contrast to the inherent nature of humans, as world-forming (*welthildend*) creatures who are capable of apprehending a being *as* such, as a part of a world comprised of a set of beings interlinked by a network of meanings[51]. An animal, Heidegger claims, does not exist, it merely lives; it sees, but does not observe[52], and this failure to objectify its environment and conceptually convert it into something abstract is precisely what traps an animal in its *Umwelt*[53], barring it from action in the larger world (*Welt*). Put simply then, the concept of *Umwelt* represents the idea that animals are

may cease to be an agent at all: "…a robot with only visual censors in an environment without light is not an agent…"(1996, 25).
[47] Firenze, 2019, 40.
[48] von Uexküll, 1957, 24.
[49] Firenze, 2019, 40.
[50] Heidegger, 1995, 225.
[51] Ibid, p. 264.
[52] Ibid., p. 210.
[53] Firenze, 2019, 44.

inserted in and enveloped by their environment, as an extension of their body without which they could not live, or could not act.

In the case of artificial agents, we can characterize their purpose-oriented ontology as one that corresponds to a specific *Umwelt*, denoting both the perceptual capacities which allow the agent to perceive its environment in a specific way (through 'marks of significance' provided by the designer), and the practical capacities that these marks of significance afford[54]. We can use the example of an autonomous vehicle to illustrate this claim: an autonomous vehicle's purpose-oriented ontology directly relates to its design objectives: to drive efficiently in real-life traffic environments. This includes the perceptual capacity to apprehend marks of significance in its environment (stop signs, road markings, pedestrians, etc.) and also the practical capacity to *act* in accordance to these marks of significance (a stop-sign provides the action of 'stopping', a solid double lane marking forbids the changing of lanes). Without these capacities, the vehicle would not meet its design objectives, since it would not be able to drive efficiently. The autonomous vehicle, in this way, is enveloped by its environment, and could not be an efficient agent if placed outside of its ideal operational context (say, in a grass field, or in a shopping mall). Its *Umwelt* is limited to urban, peri-urban and perhaps highway and rural traffic environments, and these contexts together prescribe and circumscribe the practical behavior of the autonomous vehicle. Significantly, the autonomous vehicle does not need to apprehend a pedestrian as such in order to act in certain ways towards it, it does not need to recognize a human being as such, and represent this concept abstractly, in order to exercise its practical agency[55]. Instead, it is sufficient that the vehicle correctly identify the phenomenal signature of a pedestrian, through a process of object classification.

[54] Coeckelbergh, 2011.

[55] This lack of conceptual (or metaphysical) depth will be pivotal in our investigations of the moral behavior of artificial agents in the second section of this thesis. For one thing, it is clear that common-sense morality often takes for granted a certain 'ethically thick' apprehension of moral subjects by human moral reasoners (conceiving of them as subjects with intrinsic value or rights, for example). For another, many moral theories leave substantive deliberative space for this type of apprehension, for instance when considering what types of principles would not be objected to by the parties affected by the moral decision at hand (Scanlon, 1998). We will try to capture the comparatively shallow type of apprehension alluded to here with the concept of an 'affordance' in chapter 6.

Another, closely related aspect of an artificial agent's purpose-oriented ontology is its *world knowledge*: the types of facts, social circumstances, rules and truths which must be implemented into an artificial agent in order for them to efficiently perform their design objectives[56]. Ostensibly, human agents, acting in the world (*Welt*) are privy to an exorbitant amount of facts, only some of which are useful for decision-making and practical behavior in a given context. When making a sandwich for instance, human agents might know that butter is easier to spread when soft, that a 'sandwich' requires at least two slices of bread, that sardines and peanut butter do not go well together, that people from Phoenix are called Phoenicians, and that Donald Trump is the president of the United States of America—only some of which are ostensibly useful to the task at hand. Artificial agents, on the other hand, are what Daniel Dennett calls 'Potemkin villages', "…cleverly constructed facades, like cinema sets"[57]. This is directly due to an economically-grounded limitation on just how much information needs to be programmed into any given AA, since "The actual filling-in of details of AI programs is time-consuming, costly work, so…only those surfaces of the phenomenon that are like[ly] to be probed or observed are represented"[58]. Here again then, the concept of the *Umwelt* proves useful, since the environment of a given AA, including the agents and objects which act within it, may require that specific types of world knowledge be implemented into the machine for its design objectives to be met. This necessary world knowledge is not, in turn, knowledge of the *world* at large—as humans might have—but instead the relevant facts that pertain to an AA's *Umwelt*. An autonomous vehicle may require knowledge of specific facts about its environment, for instance, that elderly people are more likely to die as a result of a collision, but it does not need to know that sardines and peanut butter don't mix.

---

[56] Dennett, 1998b, 7. (Newell, 1982) has made a useful distinction surrounding the knowledge level of an artificial agent, separating the problem at the semantic level of what information to program (world knowledge as we have defined it above) from knowledge on a syntactic level (what system, format, structure or mechanism to implement in order to make use of this knowledge). We will leave this distinction alone for our ontological concerns here, but will encroach upon it in part II through our distinction between a computational and theoretical decision procedure in artificial morality.

[57] Dennett, 1998a, 16.

[58] Ibid, p, 16. The discerning eye might detect that the type of AI program that Dennett addresses here is likely a top-down, GOFAI style of decisional architecture, one which requires that an engineer manually code all of the world knowledge into the AA's software, rather than have the AA learn or infer connections and world knowledge through a bottom-up, stochastic approach. Even if this is true, many of the artificial agents which we will address across this thesis have some elements of top-down programming, and thus the concept of world knowledge retains its relevance for our analysis. We will address different programming styles at the end of this chapter, and again in chapter IV.

Regrouping these ideas, we can use the concepts of *Umwelt* and world knowledge to rework the 5th definition of artificial agents—introduced previously, and repeated below—in order to erase the apparent contradiction between 'design objectives' and the AA's ability to 'take the initiative in order to fulfil [its] own goals':

> **Definition 5:** "Intelligent agents are those that are capable of flexible action in order to meet their design objectives, where flexibility includes the following properties: (1) Reactivity: the ability to perceive their environment, respond to changes that occur in it, and possibly to learn how best to adopt those changes. (2) Proactiveness: the ability to take the initiative in order to fulfil their own goals. (3) Sociability: the ability to interact with other agents or humans"[59].

Taking things in order, we can rephrase the first sentence as follows: 'artificial agents are technological artefacts that are capable of flexible action within a specific *Umwelt* in order to meet their design objectives'. With this, we have added an emphasis on the limited practical environment of any artificial agent, a 'soap bubble' that delineates the boundaries of an AA's efficient action. The addition of the concept of an *Umwelt* affords a substantive characterization of an environment, one which includes a perceptual realm and a practical realm, which together combine to afford certain actions to an AA, which it can identify through 'marks of significance'. We have also specified that artificial agents are technological artefacts, non-natural entities that have been built in alignment with human intentions. Next, the condition of 'reactivity' can denote the intelligent behavior of an artificial agent in function of the marks of significance within its environment, which themselves may be bolstered by certain types of world knowledge, pertaining directly to the inherent characteristics of its *Umwelt*. To capture this, we can rephrase the second sentence accordingly: 'Reactivity: the ability to perceive marks of significance, respond to changes that occur between them, and possibly to learn how best to adopt these changes, mobilizing appropriate world knowledge'[60].

---

[59] Dignum, 2019, 10.

[60] One might wonder why an artificial agent should not simply respond to changes in its *Umwelt*, and not to changes across marks of significance. This choice is one of conceptual subtlety. An artificial agent, like an animal, only has knowledge of its *Umwelt through* its perception of marks of significance, and the relevant world knowledge that pertains to them. The marks of significance themselves (and the adiaphoric

Proactivity (2), in turn requires substantial amendment. Recall first that an *Umwelt*'s marks of significance can be seen to afford (or forbid) certain actions to the agent, in other words, an agent's behavior is *determined* by its *Umwelt*'s marks of significance. This appears to require that the proactiveness of an artificial agent be somehow tethered to the marks of significance it can perceive. Here, we have an opportunity to make use of the concept of autonomy, understood as a lack of direct human supervision or control. Our first option is to say that an artificial agent's fulfilment of its own goals is entirely determined by the marks of significance within its environment. This would mean that the marks of significance provide the total set of action options available to the agent. This first, rather strong condition is inappropriate for two reasons. First, this option appears to thwart many of the intuitions that other definitions attempted to capture, notably the concept of proactivity presented in definition (4): "pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative"[61]. Crudely put, the intuition here is that artificial agents must be something more than thermostats or landmines, they should not act *exclusively* in response to their environment. Thus, it cannot be that the total set of action options or opportunities is determined by an agent's environment, since this fails to give a satisfactory account of artificial agents as *actors*, and not simply passive *reactors*.

The second reason for which this exclusive form of environmental determinism is unsatisfactory follows from the first. It is quite clearly a condition of an AA's purpose-oriented ontology that it pursue a specific *purpose*, and not simply react when that specific purpose avails itself within an agent's *Umwelt*. In other words, it must exhibit goal-oriented behavior, and pursue these goals *through* its context of action. The scope of these goals is not limitless, since it is quite clearly limited by the type of *Umwelt* in which the agent acts. For instance, an autonomous vehicle cannot pursue the goal of making a sandwich, or bombing military targets[62]. Thus, the pursuit of

---

or non-significant features it can ignore) then comprise the total set of facts and features that are epistemologically available to the agent. Thus, any significant change will happen across the changes in an *Umwelt*'s marks of significance, and not the *Umwelt* considered generally.

[61] Woolbridge & Jennings, 1995, 2.

[62] Here, we are considering this point ontologically. However, the way in which the pursuit of goals is designed by human agents matters greatly for the type of moral behavior it can pursue, and the type of

an agent's goals is at the very least upper-bounded by the *type* of environment in which it acts, and for which it is designed to act. It is not, however, reducible to the marks of significance of the *Umwelt*, since this would seem to forbid any meaningful type of goal-oriented behavior *simpliciter*.

We are now left with two conceptual limitations: on one hand, the proactiveness of an artificial agent (the pursuit of its own goals) is limited by the *type* of *Umwelt* in which the agent is designed to act, but on the other, it is clearly not *reducible* to the marks of significance of this *Umwelt*. Faced with this, we could say that the artificial agent's proactiveness can be characterized in the following way: the agent's *goals* are selected internally, originating *within* the agent, but the *means* by which it achieves its goal are entirely determined by the marks of significance of its *Umwelt*. In other words, the 'what' of the agent's actions is self-determined[63], but the 'how' of the agent's actions is determined by its environment. Intuitively, this fits with a common sense understanding of what most artificial agents are: a robot vacuum cleaner proactively *searches* for dirty floors to clean (its goal), rather than wait for the bit of floor underneath it to become dusty, or for some human to turn it on. Still, it cleans the floor (achieving its goal), in function of the marks of significance (the means) of its environment (chairs, stairs, cats, charging stations, etc.). Accordingly, we can modify the definition of proactiveness given in definition (5) in the following way: 'proactiveness: the ability to pursue internal goals in a way that is responsive to the marks of significance of its *Umwelt*'. The autonomy of an artificial agent then, denotes the scope of these internal goals, and the way in which it responds to the means, or marks of significance of its environment, since it is precisely in the exercise of this capacity that it acts without direct human supervision or control.

Finally, we must contend with sociability: "the ability to interact with other agents or humans"[64]. Ostensibly, other agents and humans feature among the marks of significance of virtually every *Umwelt* in which an artificial agent could plausibly be deployed, and thus a

autonomy it can be seen to 'possess', as we shall see in the next section of this chapter, as well as chapter II.

[63] We will be careful to avoid confusing self-determination as it is used here with self-determination as it applies to humans, often denoting some substantive form of independence or moral autonomy. This is not the case, as we shall see throughout Part I of this thesis.

[64] Dignum, 2019, 10.

reductive view would hold that the behavior described by sociability could be adequately covered by our definition of proactivity. However, what is often implied by sociability is something more than simple recognition or responsiveness on the part of the artificial agent. Instead, sociability often implies that the artificial agent can *express* or render intelligible its internal states, goals, or processes to surrounding agents, a notion captured by popular design values such as explainability, transparency, and accountability[65]. This is especially true of social robots who are designed to interact in a conversational way with human beings[66]. Accordingly, the conditions of sociability are more stringent for certain artificial agents than for others. Autonomous vehicles, for instance, have very different sociability conditions: they must be interactive with the users of the vehicle (dashboard interaction for user input, and perhaps a 'why did you do that'[67] button), but also, they must adequately communicate their intentions to the surrounding traffic environment (signaling, brake-lights, pedestrian communication[68]). For our purposes here then, we might generally say that sociability entails the artificial agent's ability to communicate its internal goals or intentions, and correspondingly, requires a sensitivity to the goals and intentions of others. A bit more formally then, we will provide this definition of sociability: 'the ability to communicate internal goals to the *Umwelt*, and the ability to apprehend the goals and intentions of its marks of significance. Taken together then, our analysis of an artificial agent can be defined as follows:

> Artificial agents are technological artefacts that are capable of flexible action within a specific *Umwelt* in order to meet their design objectives, where

[65] Dignum, 2019 : Jobin et al., 2019.

[66] In effect, it is relatively common practice in social robotics to hold that (emotional) expressivity is a desirable, if not necessary feature of robotic behavior, whether this is in order to foster user trust and cooperation, or to increase transparency and explainability. One of the flagship studies to this effect is likely MIT's humanoid robot 'Kismet', headed by Cynthia Brezeal in the 1990s. Interestingly, the ethical appraisal of emotional expressivity resembles something of a mixed bag in the literature, where some authors see it as necessary if not foundational to moral behavior (Devillers, 2017 : McDermott, 2008), while others maintain that the simulation of emotions, approval or kinship with human agents is in itself manipulative or morally problematic (Turkle, 2005 : 2017 : Borenstein & Arkin, 2016).

[67] IEEE Global Initiative, 2016, 20.

[68] An autonomous vehicle's ability to express its intentions and to negotiate with other road users is both a popular and tricky issue in the literature (Clamann, 2017 : Mahadevan et al., 2018 : Rouchitsas & Alm, 2019), owing mainly to lack of human exposure to the technology, and certain ontological features such as a silent electric motor, or the inability to communicate through eye contact or hand gestures. Car manufacturers have proposed various communication remedies, including projecting text onto the pavement in the vehicle's headway, or various sounds and light signals. Google has even adapted the aesthetic of their cars to mimic an inoffensive human face, prompting one tech journalist to describe them as 'adorable Skynet Marshmallow Bumper Bots' (Hsu, 2014).

> flexibility includes the following properties: (1) Reactivity: the ability to perceive marks of significance, respond to changes that occur between them, and possibly to learn how best to adopt these changes, mobilizing appropriate world knowledge. (2) Proactivity: the ability to pursue internal goals in a way that is responsive to the marks of significance of its *Umwelt.* (3) Sociability: the ability to communicate internal goals to the *Umwelt*, and the ability to apprehend the goals and intentions of its marks of significance.

This definition, while general, nevertheless carries a few underlying assumptions that we would do well to flesh out here. First, it is a restrictive definition, which implicitly aims at the definition of more complex types of artificial agents. It excludes then, what we have called *minimal* artificial agents such as thermostats and landmines, which cannot be seen to possess any meaningful form of proactivity or sociability. Secondly, it remains agnostic regarding the roles of human agents in the identification of a) an AA's design objectives, b) the definition of an AA's internal goals, and c) the ways by which an AA communicates with its *Umwelt*, and how it takes into consideration the goals and intentions of the human agents operating within it. The principle reason for this agnosticism is a will to avoid any normative positioning on the concept of artificial agents at this early stage. In effect, the job of defining the role of human intervention and oversight in an artificial agent often carries with it certain moral trade-offs and considerations which in turn may affect not only the type of artificial morality that can be appropriate to implement, but deeper questions concerning the moral status of the artificial agent itself. For instance, if the internal goals of an AA are self-determined in the conceptually thick sense of the term—connoting subjective preferences, moral autonomy, or consciousness akin to that of human agents—then it potentially provides the grounds for the claim that artificial agents are true moral agents, entities who are owed rights and moral obligations by the human agents with which they interact[69]. Alternatively, if the internal goals of an AA are entirely determined by a human programmer, user, or a combination of the two, then this may provide grounds for the claim that artificial agents are incapable of moral agency *simpliciter*[70], or that any ascription of moral responsibility to an

[69] Levy, 2007: Bringsjord, 2008: Himma, 2009: Sullins, 2009: Irrgang, 2006: Drozdeck, 1992: Dennett, 1998a.

[70] This in turn may provide grounds for forbidding artificial agents from implementation in contexts of moral salience (Dignum, 2019) or forbidding any meaningful implementation of artificial morality (Grinbaum, 2018).

artificial agent is dangerous or misplaced[71]. Ultimately then, "…the moral status of robots and other AI systems is a choice, not a necessity"[72], and this choice is better left to later discussions.

Finally, our definition of artificial agents is open to the accusation of a favoritism of sorts between the two principal meta-perspectives on the ontology of artificial agents; what we have in passing described as the difference between practical and experimental artificial agents. Our definition vaguely favors the former type of artificial agent, which hails from what Virginia Dignum has called the 'engineering perspective' of artificial agents, "…which posits that the goal of AI is to solve real-world problems by building systems that exhibit intelligent behavior"[73]. This view is, to various degrees, in conflict with the opposing view, what she calls the 'scientific perspective', which aims to identify which kind of computational mechanisms are needed for modelling intelligent behavior[74]. A similar distinction is made by Russell & Norvig in their classification of intelligent systems, of which there are four kinds: a) systems that think like humans where the focus is on cognitive modelling, b) systems that act like humans, with focus on simulating human activity, c) systems that think rationally by using logic-based approaches to model uncertainty and deal with complexity, and d) systems that act rationally, where the focus is on agents that maximize the expected value of their performance in their environment[75]. Given that our definition of artificial agents quite clearly presupposes action in an *Umwelt*, and thus *acting* artificial agents, we have favored the categories (b) and (d). On a more normative bend, Johnson & Miller have characterized the difference between practically oriented and the experimentally oriented artificial agents as the difference between what they call the 'computers in society' and the 'computational modelers' approach to artificial agents, respectively[76]. As they see it, computational modelers "…have a stake in using computation as a (if not *the*) foundation of a body of knowledge that brings insight to a wide range of areas…and have a stake in the value of computation as a model"[77], while the computers in society group "…has a two-part agenda: show that technology is an important component of morality but also show that technology is under

[71] Johnson & Powers, 2008: Johnson & Miller, 2008: Bryson & Kime, 2011: Bryson, 2010.
[72] Bryson, 2018, 17.
[73] Dignum, 2019, 11.
[74] Ibid., p. 11.
[75] Russel & Norvig, 2013.
[76] Johnson & Miller, 2008.
[77] Ibid, p. 126.

human control"[78]. Their position (computers in society) is but one example of the types of normative extrapolations that can arise from a practically-oriented vision of artificial agents, whose purpose is to assist, replace or emulate human roles and behavior in real-world contexts. Even if we endorse this practical view of artificial agents, as these authors have, it is important to remark that this does not necessarily preclude any involvement of experimental research into the design of artificial agents themselves, especially when it concerns questions of artificial morality[79]. However, as far as our baseline definition of artificial agents is concerned, it is clearly one which aims to capture the necessary characteristics of practical, acting and worldly artificial agents.

2. *Designing Artificial Agents in an Umwelt*

If we adopt the practical view of the definition of artificial agents—that their purpose is to assist, replace or emulate human roles and behavior in a given real-world context—we should not be surprised to find that this practicality features in many of the design decisions one could make about a given artificial agent's internal and external features. Indeed, the notion of purpose-oriented ontology goes a long way in describing what kind of artificial agent is *ideal* for a real-world context. One popular way to decipher the ideal artificial agent for a given environment is through what is called a PEAS classification[80], where the designer identifies a priori the **P**erformance measure, the **E**nvironment, the **A**ctuators, and the **S**ensors required for optimal action in a given 'task environment', which we have called an *Umwelt*.

Generally speaking, the performance measure defines the *criterion of success* of an agent's actions, and is typically thought of as a set of desirable *environment states*, rather than a set of

[78] Ibid., p. 127. "Artificial agents will be understood to be human constructions under human control. They would always be understood to be 'tethered' to humans in the sense that they are the products of human invention, are deployed by humans for human purposes, operate in contexts maintained by humans, and cannot function without some degree of human control (even though that control may be distant in time and space)". (2008, 128). A similar distinction is made in a subsequent paper between a computational artefact and an AI system, where AI systems carry many of the normative features of practically-oriented artificial agents as have just been described above (Johnson & Verdicchio, 2017).
[79] Indeed, as we shall see in chapter IV, there exists a host of so-called 'bottom-up' approaches to artificial morality which mobilize sophisticated computational methods to decipher, intuit or learn moral behavior from human data, a process which is not always tethered to a specific task or *Umwelt*.
[80] Russel & Norvig, 2013, 40.

desirable behaviors[81]. For a robot vacuum cleaner, for instance, the performance measure would reward a clean floor, and the measure for a meat-rendering robot might be the percentage of meats in the correct bins. The notion of a performance measure in turn, shares an intimate relationship with the concept of *rationality*; a rational artificial agent, throughout its practical behavior, should seek to *maximize its expected performance* in its environment, as defined by the performance measure[82]. Thus a meat-rendering robot is rational if it succeeds in maximizing the percentage of meat in the appropriate bins, and irrational if it fails to do so, by say, placing the meat in the wrong bins, or by sorting the meat but then dumping the bins out onto the floor. Two points, however, must be clarified immediately. First, the concept of rationality in a computational or engineering-oriented sense does not *necessarily* bear any relation to the concept of rationality as it is typically employed in the fields of philosophy, economics, decision theory or game theory. Rationality here is defined as the maximization of an artificial agent's performance measure, which is defined by the designer (in function of the task environment), not any overarching moral theory or decision-theoretic strategy[83]. A second point is that the definition of the performance measure is not always a straightforward task; if a 'clean floor' is the criterion of success of a robot vacuum cleaner, this tells us little about how this ought to be achieved by the agent. Should the robot clean vigorously and then lie in wait until the floor becomes dirty? Or should it constantly survey the floor, cleaning small portions throughout the day. Both options satisfy the goal of a 'clean floor', but without a detailed specification of the ideal way in which an agent approaches its task, it risks irrational or undesirable behavior[84].

---

[81] "As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave" Ibid, p. 37.

[82] Russel & Norvig define a rational agent in this way: "…for each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has". (2013, 38). The percept sequence, in turn, is comprised of individual percepts (perceptual input at any given time). A percept sequence then, is the complete history of an artificial agent's perceptual input.

[83] This being said, nothing precludes a designer from taking inspiration from these fields, as is the case with what are called *normative expert systems* (Horowitz & Heidermann, 1986) which act rationally according to the laws of decision theory, whether or not this aligns with analogous human decision processes (Russel & Norvig, 2013, 26). As we shall see, various proposals for artificial morality also seek to define ideal moral behavior in terms of 'rationality' understood in the philosophical sense, with a goal of producing cooperative behavior among self-interested organisms in repeated cooperation games (Leben, 2018, 50). This is one, relatively universal way to tackle the problem of moral behavior in artificial agents, but we will attempt to show that there remain other, perhaps more acceptable options on offer.

[84] Russel & Norvig, 37.

Beyond the type of task an agent is destined to perform, and the ideal way in which it should perform it, the environment of an artificial agent also affords a great wealth of information about the types of hardware and software with which an artificial agent will need to be equipped. A typification of an agent's environment can pass through many terms and distinctions, but for our purposes here, we will provide and explain the main aspects of most environments in which artificial agents are typically destined to act. For the brunt of artificial agents then, their task environments are[85]:

Partially observable: an AA does not have access to the complete state of its *Umwelt*, or at least the complete state of all of its marks of significance at each point in time. For instance, an autonomous vehicle's vision may be blocked by a large truck directly in front of it, and thus it is unable to perceive the totality of its surroundings. Or, an autonomous vehicle is unable to know what other drivers in the traffic environment are thinking.

Multi-agent: an AA acts within an *Umwelt* in which multiple agents are present, who may be seen to react to the AA's actions and choices. An autonomous vehicle, for instance, must act and interact within a traffic environment, where other human drivers, pedestrians, cyclists and animals may be present. The environment can be cooperative (i.e. all agents work to minimize the risk of traffic accidents), or competitive, as in the case of a game of chess.

Stochastic: the next state of an AA's *Umwelt* is not completely determined by the artificial agent's current state, and the action executed by the artificial agent. For instance, the next state of the traffic environment cannot be solely determined by the actions of an autonomous vehicle, since the traffic environment in general depends on the actions of multiple agents, and unforeseen events can happen, such as brake failure or a blown-out tire.

Sequential: the current decision of an AA affects all future decisions that the AA can make. An AA's action, in other words, depends on the actions it has previously chosen to pursue. An autonomous vehicle does not 'reset' at a specific location relative to its environment at the outset of each of its decisions, it makes decisions (i.e. to change lanes) that affect the decisions it can make in the future.

Dynamic: The *Umwelt* of an artificial agent can evolve while the artificial agent deliberates on which action to pursue. An autonomous vehicle makes decisions *as* it is driving, the traffic environment which surrounds it does not courteously grind to a halt while the vehicle decides what to do.

[85] We borrow these distinctions from Russel & Norvig's discussion of task environments (2013, 42-46), adding our vocabulary from the previous section.

Continuous: Pertaining to both the evolution of an artificial agent's *Umwelt*, and the passage of time within it, a continuous environment has an infinite number of states, and the position of its marks of significance "..sweep through a range of values and do so smoothly over time"[86]. Autonomous vehicle driving is a continuous-state and continuous-time problem, since the other vehicles move smoothly across the environment over time, but so too do the vehicle's own movements (steering angles, velocity, etc.).

The typification of an artificial agent's *Umwelt* has ramifications for both the internal and external aspects of the AA's design. Internally, it affects the structure of an artificial agent, specifically the type of *agent program* that will be implemented into the artificial agent. Simply put, the agent program is the method by which an artificial agent moves from percepts (the perception of its environment) to action. It can, for instance, be mapped by *condition-action rules* (IF car in front is braking, THEN initiate braking), *model-based programs* (where the AA has an internal model of its *Umwelt*, in which it can 'test out' potential actions), *goal-based programs* (in which the AA has not only an internal model but a goal which it attempts to achieve when deliberating about potential actions), or utility-based programs (where the AA maximizes expected utility across its action options in the internal model of its *Umwelt*, where expected utility is computed by averaging over all possible outcome states, weighted by the probability of the outcome)[87].

The choice of agent program is also relative to the specification of the performance measure, which is itself a function of the agent's *Umwelt*. For instance, in the case of autonomous vehicles, there are many different values for which we would find it desirable to maximize: safety, efficiency, legality or passenger comfort being some prominent examples. This may lead designers to adopt a utility-based program to map percepts to actions, since it enables the maximization of multiple values across this agent's actions. Furthermore, we may choose this over other potential agent programs in regards to other features of the vehicle's environment, for instance, the fact that the vehicle's environment is partially observable and multi-agential, requires that the mapping from percepts to actions be probabilistic: the vehicle cannot be certain of its location, the location of marks of significance, or the future consequences of its actions. In general then, the proper

[86] Russel & Norvig, 2013, 44.
[87] Ibid, p. 54.

typification of an artificial agent's environment is paramount to the details of its purpose-oriented ontology, from what it can be seen to maximize (or the type of 'rationality' it exhibits), right down to how it moves from perception to action.

Externally, it is relatively straight-forward that an *Umwelt* goes a long way in deciding an artificial agent's physical aspects, or the Actuators and Sensors of its PEAS classification. Actuators should be conceived, roughly, as everything and anything an AA needs to act and respond to its environment. Sensors, on the other hand, are everything it needs to perceive it. For instance, an autonomous vehicle would likely need actuators akin to those available to human drivers: steering, accelerators, brakes, signals, horns, and displays. Similarly, it would need a whole host of sensors in order to efficiently act in a traffic environment: cameras, Lidar, speedometers, GPS, odometers, engine sensors, among others. Here again, the choice of actuators and sensors are related to the performance measure which provides the criterion of success for the artificial agent. If the performance measure of an autonomous vehicle includes passenger comfort, it might require that the vehicle be equipped with a thermostat which measures the ambient temperature inside the vehicle, or may also require haptic sensors which will inform the vehicle of the passenger's position.

3. *Classes of Artificial Agents & Ethical Design Concerns*

The *Umwelt* of an artificial agent, as we have just seen, provides the designer with a wealth of solutions to the design problems of an artificial agent. If the designer is able to typify the agent's *Umwelt* in a robust way, this can afford him insight into a) how the agent ought to act, or what it ought to achieve (its 'rationality', a function of its performance measure), b) the way the agent ought to move from perception to action (its agent program, for instance a utility-based program), and c) the hardware it needs to sense and act in its environment (its sensors and actuators). When this process is successful, it yields a highly specified artificial agent, capable only of efficient action in the *Umwelt* for which it was designed.

Nevertheless, across the myriad artificial agents to which the modern world is privy, certain general classifications can be made, based on either the type of *Umwelt* in which an agent acts, or

the type of programming style by which it operates. These distinctions are useful for our analysis in so far as they are themselves subject to moral scrutiny: some programming techniques are more sensitive to ethical design concerns than others, while different types of artificial agents can cause different types of moral damage. For clarity's sake, it is important to specify that there is a salient difference between what we have been calling *artificial morality* and these ethical design concerns. The former describes how a particular artificial agent responds to the moral value of its environment, while the latter describes how the design choices of a human agent (generally a designer or engineer) may fail to adequately respect human rights, dignity, or undermine societal welfare. In this way, a given artificial moral agent—which we will recall, is an artificial agent equipped with artificial morality—can still fail to meet the demands of various ethical design concerns, although this is hardly desirable. This distinction is often lost in more carefree debates surrounding the 'ethics of AI'.

Indeed, certain authors within the field of machine ethics have been concerned with the impact that this categorial vagueness might have on the proper pursuit of ethical AI, as one author ponders, "…is roboethics simply the search for a list of rules that any and all roboticists must follow in their work, such that all who adhere to the rules are automatically moral, and those who break them automatically immoral?"[88]. The truth is, unfortunately, more complicated than this. In effect, there are at least three senses of the term 'machine ethics' which are used interchangeably in current debate. First, machine ethics could denote an applied ethics, including philosophical studies about the ethical issues arising from the effects of the application of artificial agents on our society. Second, machine ethics could denote a moral code to which the artificial agents themselves are viewed to adhere. Finally, machine ethics could also denote the self-conscious ability of robots to perform ethical reasoning, "…to understand from a first-person perspective their choices and responsibilities, and to freely and self-consciously choose their course of action"[89]. In a similar vein, Virginia Dignum separates the field of machine ethics into a) ethics *in* design—denoting AI principles, regulatory and engineering processes, b) ethics *by* design—-ethics

[88] Abney, 2012, 37.
[89] Veruggio, 2005.

of the behavior of artificial agents, and c) ethics for designers—codes of conduct, regulatory requirements, standards and certifications[90].

Intuition points to a link between the first sense of machine ethics (applied ethics and philosophical analysis) with (a) and perhaps (c). This, broadly speaking, is the territory covered by ethical design concerns[91]. We might remark a rather substantial gap, however, between these two interpretations. This, too, must be explained. Until recent years, machine ethics, and specifically, the normative aspects of the field, were relatively sheltered from general concern and public attention. Machine ethicists, in the early years[92], inherited much of their method and discourse from sub-disciplines like science and technology studies and philosophy of technology, and married contributions from the fields of engineering and computer science, philosophy, and the social sciences[93]. However, the recent AI explosion (and the 'data boom' by which it was accompanied) has made machine ethics something of a pressing concern for mankind, and has encouraged contributions from a variety of other disciplines or 'stake-holders': private industry, legal experts, consumer interest advocates, governmental institutions, think-tanks, public policy advocates, among many, many more. This interest has manifested itself in the production of a flurry of 'AI principles', recommendations, and guidelines—84 individual proposals to be exact[94]—all of which make claims and recommend norms that can easily be said to target ethical

[90] Dignum, 2019, 6-7.

[91] On the other hand, ethics *by* design (b) and the second sense of machine ethics (the moral code to which robots adhere) is generally what is considered to fall under the purview of artificial morality, since it is precisely through these methods that an artificial agent responds to moral value. The third sense of machine ethics (rules and principles an AA selects for itself) is an entirely different matter, and typically denotes what are called 'full ethical agents' (Moor, 2006) in the literature. Humans, too, are full ethical agents, and this correlation tends to prompt the type of normative claims about the desirability of 'artificial personhood' which we briefly explored in the first section of this chapter. We will explore them again in the next.

[92] It is important to specify that these 'early years' were not all that long ago. Seminal literature began at the very least in the early 1990s, with what is today considered seminal literature arriving in the early 2000's, such as that of Floridi & Sanders (2001: 2004), van den Hoven & Weckert (2008), and Wendall & Wallach (2008).

[93] Bostrom & Yudkowsky have pointed out certain ethical lacunae in this phase of the field of machine ethics that would one day come to represent the central focus of ethical design concerns. As they put it: "Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgement of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers" (2014, 319)..

[94] Jobin et al., 2019.

design concerns[95]. An exhaustive cross analysis performed by Jobin et al. revealed the emergence of five core ethical design concerns: transparency, justice and fairness (equality), non-maleficence, responsibility (or accountability), and privacy[96].

These principles are meant to apply to AI technology in a very broad sense, anything from the design, collection and use of data sets, to the places in which AI is implemented, and how it interacts with human beings. As far as artificial agents are concerned, most of the emphasis revolves around the so-called ART principles: accountability, responsibility, and transparency[97]. We will briefly define these principles below, and then flesh them out against the backdrop of four general classifications of artificial agents: embodied, virtual, deterministic, and stochastic[98]. Our approaching ethical design concerns in this way is not accidental, as we shall see, beyond the demands of an artificial agent's *Umwelt*, these ethical design concerns likewise curtail the types of artificial agent that are appropriate to design and implement[99].

3.1 *ART Principles as Ethical Design Concerns*

Taking things in order, at its most general level, accountability "…refers to the requirement for the system to be able to explain and justify its decisions to users and other relevant actors"[100]. This first ethical design concern is therefore *techno-centric*, it involves the *artificial agent*'s capacity to explain its actions and decisions, typically in ways that are decipherable to the layman, and not only to the expert or programmer. Metaphorically, a good way to capture the concept of accountability is through a 'why did you do that? button'[101]: when an autonomous vehicle stops

[95] Of notable depth and thoroughness are: the Montreal declaration on responsible AI (2019), the IEEE's Ethically Aligned Design initiative (2018), and COMEST's report on the ethics of robotics (2017).
[96] Other prominent design principles were: beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity. (Jobin et al., 2019).
[97] Dignum, 2019, 53.
[98] As the application case of the arguments of this thesis revolves around autonomous vehicles, we will pay significantly less attention to what we will come to call *virtual* agents, even if many of the considerations we will discuss can be seen to apply to both.
[99] "...ART imposes requirements on AI systems' design and architecture that will condition the development process and the systems' architecture" (Dignum, 2019, 54). This problem will be taken up in greater length in chapter V.
[100] Dignum, 2019, 53
[101] IEEE Global Initiative, 2016, 20.

for a pedestrian who crosses the street 300 meters in front of the vehicle's position, a passenger may wonder why it chose to brake, rather than slow down or continue with caution. An accountable autonomous vehicle might reply: 'I perceived a risk to pedestrian safety, and I am programmed to privilege pedestrian safety at all times'. Intuitively, accountability shares a close relationship with user trust; through access to the 'how' and 'why' of an artificial agent's decisions, users may find them more trustworthy and reliable, even when mistakes happen[102]. On a more philosophical tangent, the choice of building *accountable* (and not *responsible*) machines is of particular importance to moral philosophy: it may help quell concerns about the moral agency of artificial agents. In effect, certain models of moral responsibility—particularly of the contractarian tradition—see moral responsibility as a capacity for answerability[103]. By providing moral reasons and justifications for their actions, artificial agents are seen to be 'responsible' for them, and they 'open themselves up' so to speak, to the evaluative attitudes of others. For reasons we will explore in the next chapter, artificial agents capable of true moral responsibility are not always an ideal design choice, in part because this tends to focus moral wrongdoing, misconduct, or harm *onto* technology and *away from* those who designed it[104]. But, by making a distinction between moral responsibility in human agents, and simply moral *accountability* in artificial agents, we may be able to account for the causal role AAs undoubtedly play in morally salient actions (and their ability to explain these actions), without falling into the trap of *attributing* moral responsibility to them[105].

[102] Nissenbaum, 1996, 27. "A culture of accountability is particularly important for a technology still struggling with standards of reliability because it means that even in cases where things go awry, we are assured of answerability".

[103] Scanlon, 2008, 189. "…someone is responsible for an action or an attitude just in case it is connected to her capacity for evaluative judgement in a way that opens her up, in principle, to demands of justification from others". Another way to put this is what Gary Watson calls 'aretaic appraisals', an evaluation of an agent's "excellence and faults—or virtues and vices—as manifested in thought and in action" (1996, 10).

[104] Johnson & Miller, 2008. This idea is sometimes captured by the concept of 'moral buffers' (Cummings, 2006).

[105] (Floridi, 2008, 54-55) makes a similar argument by analogy with the case of rescue dogs: "There is nothing wrong with identifying a dog as the source of a morally good action, hence, as an agent playing a crucial role in a moral situation and, therefore, as a moral agent…Emotionally, people may be very grateful to the animals, but for the dogs it is a game and they cannot be considered morally *responsible* for their actions. The point is that the dogs are involved in a moral game as main players and, hence, that we can rightly identify them as moral agents *accountable* for the good or evil they can cause".

Responsibility, the second ART principle, typically denotes the *human* role in the design of artificial agents and their resultant actions and decisions, which is, for the quasi-totality of authors, the appropriate loci of moral responsibility, or object for our moral attitudes concerning AI technology[106]: "As the chain of responsibility grows, means are needed to link the AI systems' decisions to their input data and to the actions of stakeholders involved in the system's decision"[107].The idea here is generally twofold: first, establish meaningful links between relevant decision-makers in an artificial agent's design process, and second, promote the awareness and transparency of these links in the public's understanding of AI systems[108]. The interest in human responsibility for AI systems is a long-standing concern within the field of technoethics and philosophy of technology[109], underscoring the basic idea that "…Robots are simply tools of various kinds, albeit very special tools, and the responsibility to ensure they behave well must always lie with human beings. In other words, we require ethical robotics (or roboticists) at least as much as we require ethical robots"[110]. Put another way, responsibility as an ethical design concern excludes two undesirable outcomes: a designer deflecting blame onto a machine for morally dubious consequences which were in his power to prevent; or the careless design of machines for which a chain of responsibility is difficult to establish.

Finally, we arrive at the final ART principle, Transparency. In some respects, both accountability and responsibility presuppose a fair degree of transparency in artificial agents. For instance, a transparent machine is one that is accountable and answerable to its human user, and this bolsters general public trust and confidence[111]. In other words, "…in general technology is trusted if it brings benefits while also being safe, well-regulated, and—when accidents happen—

[106] Attempts to attribute moral responsibility to machines are hardly ever targeted towards *current* artificial agents (Bostrom, 2014), or if they are, often pass through a phenomenological or external account and reworking of traditional conceptions of moral agency (Gunkel, 2012: Coeckelbergh, 2011). This is quite separate from accounts of robots *appearing* to exhibit moral agency and responsibility, for which there is some empirical evidence (Malle et al., 2016).
[107] Dignum, 2019, 54
[108] IEEE Global Initiative, 2016, 18.
[109] Mario Bunge's seminal essay '*Towards a Technoethics'* paints this point candidly, and implores the designers and engineers of sociotechnical systems to take responsibility for the technology they create, and in some respects, 'internalize' the potential moral externalities of their creations in their design choices (1977).
[110] Vanderelst & Winfield, 2018, 5.
[111] IEEE Global Initiative, 2016, 19.

subject to robust investigation"[112]. In detail, "Transparency indicates the capability to describe, inspect, and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and the provenance and dynamics of the data that is used and created by the system"[113]. What this often implies is that an artificial agent's decisions ought to be traceable or replicable, it should not be the case that *no* agent in a given AA's sociotechnical system is capable of explaining why it made a certain decision, and that normatively, AAs should not be designed in ways that allow this to occur. At a more general level, however, transparency has much to do with *where* an artificial agent acts, and whether the human agents in its specific *Umwelt* are aware of its existence and relative decisional authority. An interesting application case for transparency is the popular criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), used to assess a criminal defendant's likelihood of committing a further crime, or their recidivism rate[114]. Proponents of this system—which has been in use since around the year 2000—claim that the 137 features across which it analyses criminal defendants, allow it to make recidivism predictions which are more accurate and fair than typical human decision-makers[115]. Nevertheless, COMPAS has been met with harsh criticism in recent years, specifically for its apparent proclivity to bias its predictions against black offenders[116], despite the features not including any overtly racially sensitive criteria. But a deeper problem lies in the lack of transparency of the commercial software itself: Northpointe, the company to which COMPAS belongs, "…has not revealed the inner workings of their recidivism prediction algorithm…"[117] to the general public.

Transparency then, is an issue at multiple levels. First, transparency relates to the *use* of COMPAS (whether, for instance, the criminal courts that make use of the software disclose this use to the concerned parties), second, it relates to the *data* that the software uses (is this data publicly available, scrutable, and lawful?), and lastly, whether the *algorithms* that make these predictions are transparent to the public (how are these features weighted in the overall assessment

[112] Winfield/ Jirokta, 2017, 4.
[113] Dignum, 2019, 54.
[114] Dressel & Farid, 2018.
[115] Perry, 2013.
[116] Angwin et al., 2016.
[117] Dressel & Farid, 2018, 1.

of the software?)[118]. In this way, transparency is something of the mother of all ethical design concerns, and it is therefore no small wonder that it features in 73 of the 84 AI principle proposals that have been generated to date[119]. Most generally, transparency is heralded as something of an all-purpose route to minimizing the harm which could potentially be brought about through the design, implementation and use of AI systems[120].

3.2 *ART Principles and the Embodied-Virtual Distinction*

In the modern world, a designer has two options for the type of environment in which he aims to implement an artificial agent: the real world, or the virtual world. Real world artificial agents exist in physical space, and are correspondingly often called *embodied agents*, or more colloquially, robots. Artificial agents that exist in virtual space go by different names: algorithms, software bots or agents, or virtual agents being some of the most common options. Importantly, while the choice of the agent's environment has an inevitable and substantive impact on every aspect of its PEAS classification (what it ought to achieve, how it moves from perception to action, its actuators, and its sensors), neither environment is 'easier' to tackle for the designer[121]. Still, from a design perspective, there are some further challenges associated with embodied forms of artificial agents. First, embodied agents generally require *mobility*[122], or the ability to navigate a physical environment. Efficient mobility in embodied artificial agents has traditionally posed a

[118] As Diakopolous points out, this is sometimes easier said than done: "Whereas data transparency can be achieved by publishing a spreadsheet or database with an explanatory document of the scheme, transparency of an algorithm can be much more complicated, resulting in additional labor costs both in the creation of that information as well as in its consumption" (2017, 821).
[119] Jobin et al., 2019.
[120] "Primarily, transparency is presented as a way to minimize harm and improve AI" (Jobin et al., 2019, 391), "…a lack of transparency generates a high magnitude of harm…" (IEEE Global Initiative, 2016, 19).
[121] Russell & Norvig drive this point home : "In fact, what matters is not the distinction between 'real' and 'artificial' environments, but the complexity of the relationship among the behavior of the agent, the percept sequence generated by the environment, and the performance measure…software agents…exist in rich, unlimited domains. Imagine a softbot Web site operator designed to scan Internet news sources and show the interesting items to its users, while selling advertising space to generate revenue. To do well, that operator will need some natural language processing abilities, it will need to learn what each user and advertiser is interested in, and it will need to change its plans dynamically—for example, when the connection for one news source goes down or when a new one comes online. The internet is an environment whose complexity rivals that of the physical world and whose inhabitants include many artificial and human agents" (2013, 41).
[122] COMEST, 2017.

lofty challenge for engineers, from walking and ascending stairs to the manipulation of every-day objects, such as doorknobs, cups or stools. In other words, the actuators of the embodied artificial agent must enable the full range of mobility necessary to ensure that their performance measure is attainable in their environment, and if this is not possible, it may affect the types of tasks they can accomplish, or the overall quality of their agency[123].

Furthermore, and of considerable ethical salience, embodied forms of artificial agents introduce the capacity for *physical harm to human agents*, and there has been a long history of such unfortunate events occurring when humans fail to interpret and adapt to the behavior of an AA[124]. Beyond these lethal mishaps, there are also what are called *safety critical systems*[125], technology whose very purpose implies some potential for human harm, such as autonomous weapons or autonomous vehicles[126]. Virtual agents, as we have seen in the case of the COMPAS recidivism prediction agent, are not above harm and wrongdoing: through their decisions, actions and predictions, human agents may lose opportunities, experience discrimination, or generally have their interest and welfare thwarted. It is difficult, however, to maintain that they cause direct lethal harm[127]. The risk of physical harm then, is a major burden on the design of embodied artificial agents, a concern which in itself may preclude their implementation into certain spheres of human society[128].

---

[123] Some designers have risen to this challenge: Spot, the robot dog from Boston Dynamics, can run (up stairs), and resist falling over when kicked. Pepper, a popular humanoid robot assistant, may have a hard time with stairs, but is able to catch a ball in a cup.

[124] Two of the most iconic cases are the death of Robert Williams, who was killed by an industrial robotic arm in a Ford Motor Factory in 1979 (Ottawa Citizen, 1983), and Kenji Urada, who was killed in similar circumstances at a Japanese industrial plant in 1981 (United Press International, 1981). Incidentally, the subsequent legal investigations claimed that both men were insufficiently knowledgeable about the kinds of operations the robot could perform.

[125] Chatila et al., 2017, 129.

[126] Interestingly, while both systems are likely to cause harm, intentional lethal action only figures in the performance measure of autonomous weapons. Autonomous vehicles, on the other hand, may be obliged to cause harm as a result of the complexity of their environment, as is the case in unavoidable accident scenarios. Whether or not this harm can be avoided with the ideal design of an autonomous vehicle is a question we will consider in the last section of this thesis.

[127] One such lethally harmful virtual agent is Nick Bostrom's well-known Paperclip Maximizer, an artificial agent endowed with general intelligence whose careless programmers designed it to consume any and all of the world's resources to maximize the production of paperclips (Bostrom, 2003).

[128] Consider, for instance, the public outcry surrounding the implementation of so-called 'killer robots' (www.stopkillerrobots.org).

As far as ART principles are concerned, the choice between embodied and virtual artificial agents may further affect certain design decisions. Focusing on the case of embodied agents, we can use the practical example of autonomous vehicles to hash out these concerns. In terms of accountability, or the ability of an AA to provide an explanation or justification of its actions and decisions, the persons to which an autonomous vehicle may need to be held accountable often maintain closer relationships with the vehicle than is the case of virtual agents. Autonomous vehicles, in other words, carry *passengers*, whose personal autonomy is therefore delegated to the vehicle itself. It is the vehicle who decides how a given passenger arrives at their desired destination, and when an accident is unavoidable, it is the vehicle who decides how to crash[129]. In these unfortunate cases then, it is the vehicle that must account for why it ostensibly 'decided' to harm the passenger rather than the pedestrian, or in more mundane cases, why it decided to make its passenger late for work, rather than run the risk of bumping into a j-walking pedestrian. The 'why did you do that? Button' in the case of autonomous vehicles, in turn, often manifests itself in the *ethics settings* that are programmed into the vehicle[130], which are ideally visible to the passenger through, for instance, an interactive dashboard, and may even be modifiable by the passenger[131]. In this way, the passenger is aware of the normative considerations that underpin his vehicle's decisions, and is not unfairly caught by surprise when his vehicle chooses to sacrifice him rather than a criminal offender[132].

Transparency in embodied agents follows along similar lines, with the additional concern of the data that is collected and used by these systems. For instance, it is clearly an ethically salient question to know *what* and *how much* an autonomous vehicle can know about its surrounding environment, especially when this information is used to make lethal decisions[133]. Should human

[129] This of course presupposes a level 5 autonomous vehicle (SAE, 2016) capable of full autonomy in all driving contexts and conditions.
[130] Millar, 2014b: Millar et al., 2017. This term is tantamount to an autonomous vehicle's artificial morality.
[131] Whether or not a passenger should be 'allowed' to modify the ethics settings of his autonomous vehicle is a highly debated issue in the ethics of autonomous vehicles (Lin, 2014a: Millar et al., 2017: Gogol & Mueller, 2017: Barghava & Kim, 2017).
[132] In an attempt to understand society's deep-rooted intuitions about ethics settings, the MIT moral machine experiment developed an online serious game platform which indeed pitted passengers against criminals (Bonnefon et al., 2016). We will spend much time with the ramifications of this study in part II.
[133] We address this problem in chapter V.

beings be ranked according to their social status, occupation, age, weight or health? These sorts of dystopian considerations are bolstered by advances in vehicle-to-vehicle and vehicle-to-device communications, the Internet of Things, and various 'sniffing' protocols that seek to identify and track individual users as they move through a traffic environment. Generally, transparency is concerned with the type of data that is being *used* by the autonomous vehicle, but also the type of data that is being collected *by* the autonomous vehicle, and whether this information and collection is accessible, intelligible or consented to by users, the general public, and crash investigators.

Finally, the problem of responsibility has been a long-standing issue in embodied artificial agents, most especially those that are considered safety critical systems. In the case of military weapons, it is an open question whether the introduction of autonomous technology into the military chain of command will not exacerbate the incentive to war[134], generate moral buffers[135] between human decision-makers and the machines that execute these decisions, or provide various economic incentives to privilege the sanctity of expensive military equipment over human lives[136]. In general, while the military chain of command (and the chain of responsibility it underpins) is notoriously clear-cut, the introduction of artificial agents may not help, but rather hinder meaningful accountability on the part of human decision-makers. These concerns seem to point to the idea that the responsible implementation of autonomous weapons may not be possible *in toto*, despite convincing contrarian positions[137]. The case of autonomous vehicles, in turn, is slightly different. *In praesentia*, the absence of clear and appropriate legislation as to the locus of liability for an autonomous vehicle's actions, be they lethal or simply damaging, has created 'gaps' in the public's vision of who may be held responsible for an AV's decision[138]. Should responsibility for damage caused be synthetically linked to a proximate human agent? Who should shoulder the blame, the original equipment manufacturer, or the passenger or owner of the vehicle[139]? As autonomous vehicles encroach on more and more of the world's—and especially America's—

[134] This is sometimes characterized as the risk of a 'push-button war' (Sparrow, 2007).
[135] Cummings, 2006, 35. John Sullins also calls this the 'distance problem' (2006, 28).
[136] Anderson & Waxman, 2012.
[137] Most notably, Arkin, 2009.
[138] While much of the literature focuses on so-called liability or accountability 'gaps' in autonomous weapons (Docherty, 2015 : Gunkel, 2017 : McFarland & McCormack, 2014); in the literature on autonomous vehicles, a corresponding 'retribution gap', relating to a victim's reclamations of justice for the vehicle's harmful actions, has been addressed by (Danaher, 2016).
[139] Gurney, 2015.

roads, 'driverless accidents' like these are an ever more frequent affair[140]. In this way, the demands of responsibility are a true ethical design concern. Here again, the failure to properly account for a chain of responsibility may result not only in ethically dubious implementation, but also a lack of public trust and acceptability.

3.3 *ART Principles and the Deterministic-Stochastic Distinction*

While the physical distinction between embodied and virtual artificial agents is clearly pertinent to the types of behavior these agents can perform, and to a certain degree, the types of environment that are open to responsible design and implementation, a far more prevalent distinction can be made concerning the types of system design, or programming, on which an artificial agent is based. There are two general categories within which a given artificial agent could fall: *deterministic*—sometimes called top-down programming or 'expert systems'—and *stochastic*—-also called, probabilistic, bottom-up, machine learning, or 'learning from data' approaches. Most of the buzz concerning the 'AI boom' of recent years concerns the latter, machine learning type of technology, since it is this type that has seen a resurgence in popularity with the advent of the internet, and the massive amounts of accessible data which it provides[141]. Nevertheless, deterministic expert systems remain quite popular in certain areas of engineering and robotics, all the more so in areas where it is desirable that an artificial agent abide by strict rules and constraints[142]. We can see why if we explore the definitions of these two approaches in more detail.

---

[140] Consider, for example, the object classification error that lead a Tesla Model S to collide with a parked fire truck in 2018. In a telling response, the American National Transportation Safety Board issued a statement maintaining "…The probable cause of the Culver City, California, rear-end crash was the Tesla driver's lack of response to the stationary fire truck in his travel lane, due to inattention and over reliance on the vehicle's advanced driver system; the Tesla's Autopilot design, which permitted the driver to disengage from the driving task; and the driver's use of the system in ways inconsistent with guidance warnings from the manufacturer"(LA Times, 2019).

[141] Still, many of the connectivist theories and models which underpin these types of approaches have existed for over half a century, and machine learning methods themselves since the 1980s (Kearns & Roth, 2019: Russel & Norvig, 2013: Brooks, 2003).

[142] As we shall see, the vast majority of proposals for artificial morality, especially those that seek to emulate specific moral theories, rely on a deterministic structure to select permissible and forbidden actions, precisely because, like humans, one popular conception of morality has to do with strict adherence to rules,

The deterministic, or expert system approach, most generally, "…aims to emulate the principles used by human experts", wherein "programmers sit down with human domain experts to understand the criteria used to make decisions, and then translate these rules into software code"[143]. This often yields "…basically pre-programmed and essentially deterministic"[144] behavior. In other words, this ensures that the resultant behavior of the artificial agent is highly *predictable*, and strictly follows the rules and constraints set forth by the human designer. Historically, expert systems are the most loyal to the early aims of artificial intelligence, and often share many characteristics with what is rather fondly called Good Old-Fashioned Artificial Intelligence (GOFAI). Indeed, the enthusiasm for expert systems dates back to the now infamous Dartmouth summer workshop on artificial intelligence in 1956, what some consider to be the birthplace of the field of AI. We can see this connection in the workshop's proposal: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[145]. Clearly, the job of 'precisely describing' various features of intelligence was a human one, and while the admirable universality of these imagined systems did not carry over into modern day expert-systems, the preference for human-given, top-down programming remains[146].

On the other side of the spectrum, there are bottom-up or machine learning approaches. These approaches, in turn, "…are often based on stochastic or statistical methods to parse, compare, and extrapolate patterns from a set of data. In most cases, these algorithms require data, and lots of it. This data is used to train algorithms, often over a long period of time, until they are able to correctly identify patterns and apply knowledge to similar situations"[147]. In other words, if expert systems involve the explicit design of a decision procedure that mimics human reasoning, machine learning approaches mainly leave the decision procedure up to the machine itself; using

---

the evaluation of explicit principles or maxims, and the respect of duties and obligations across an agent's behavior.

[143] Dignum, 2019, 22.

[144] COMEST, 2017, 4.

[145] McCarthy et al., 2006.

[146] In other words, today's expert systems do not aim at general intelligence—or as Daniel Dennett puts it, the 'Master Program'—as the fathers of AI had imagined it. Instead, they aim to efficiently describe optimal behavior in a limited action context, or *Umwelt*.

[147] Dignum, 2019, 26.

massive sets of training data to 'teach' the machine until it yields the desirable output. Machine learning algorithms are at work in many corners of modern society, most especially in natural language processing and facial or image recognition and classifications, jobs that have been notoriously hard to accomplish by expert systems. This tradition is owed, at least in part, to the work of Alan Turing in the early 1950s, who introduced the notion of the Child Programme : "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulated the child's?"[148]. Another early precursor was Marvin Minsky of the 'connectivist' school, who along with Dean Edmonds, built the first neural network computer in 1950[149].

Even if both of these approaches start at opposite ends, this does not preclude them from being equally viable solutions to every-day engineering problems. One example would be a hypothetical 'Admissions Bot' that decides whether applicants are admitted into a university:

> The program is tasked with reviewing each applicant's file and making an admission decision based on a student's SAT score, grade-point average and a numerical score assigned to the difficulty of his or her high school's curriculum. The first computer program applies hard rules—multiply the SAT score by 10, the grade-point average by 6, and then adjust based on difficulty by multiplying by the high school difficulty score. Then, rank the scores and the students in the top 10% of the scores are admitted. The second computer program is given the same data about the candidates for admission—SAT score, grade-point average, high school difficulty—but is also given historical admissions decisions as well as the corresponding SAT, grade-point, and difficulty scores. Because the second computer program is not given hard and fast rules, it must devise its own way of determining which students to admit and which to reject, based on its knowledge of past data[150].

The first computer program is an expert system, and the second is clearly a stochastic or machine learning program. Superficially, we may not have any decisive reason to choose one program over the other, as both seem to accomplish the desired task, in accordance with the performance measure set forth by the designer. However, in the literature on ethical design

[148] Turing, 2004.
[149] Russel & Norvig, 2013, 16.
[150] Bathaee, 2017, 898-899.

concerns, there are clearly some reasons to prefer the first program to the second, and we will work through them in alignment with the principles of accountability, transparency and responsibility.

Many of the problems posed by machine learning approaches to artificial agents relate directly to what is called the *black box problem*, "…the difficulty for the system to provide a suitable explanation for how it arrived at an answer…"[151], classification, prediction, or decision[152]. In the case of the Admissions Bot, for instance, while we have good knowledge of the information that forms the *basis* of the machine's decision (the history of college admissions, the SAT score of applicants, their grade-point average, etc.) we do not know precisely how the machine *made* the decision, specifically, we do not know which features mattered most to the machine, and how they compared to others. Black boxes like these are therefore *opaque* to human understanding, and only some types of stochastic programs are responsive to reverse engineering; the process of articulating the specifications of a system through rigorous examination to unearth a model of how that system works[153]. By contrast, expert systems are necessarily transparent to human understanding, since, simply put, it is precisely this understanding that designed the specific decision procedures in the first place[154]. Furthermore, they are much more amenable to reverse-engineering techniques, especially if the source code is available to the inspector. Thus, both in terms of accountability and transparency, stochastic or machine learning approaches are burdened with additional hurdles if they are to meet the conditions of ethical design, while expert systems are highly accountable and transparent, and in this way, more desirable.

---

[151] Adadi & Berrada, 2018, 52145.

[152] There are two separate concepts that contribute to the problem of a black box: complexity and dimensionality. Complexity is generally reserved for neural network-type models, and points to the complexity of connections between multi-layered networks of neurons, none of which individually hold any specific 'part' of the decision-making process. Dimensionality, in turn, is typically reserved for algorithms such as support vector machines (SVM) which treat many variables at once in the decision-making process. Roughly, each variable counts as a dimension, and the more dimensions there are, the harder it is for a human being to conceive of—let alone understand or explain—the decision-making process.

[153] Eilam, 2005. Of particular resistance to reverse engineering techniques are so-called neural networks, where the process rivals that of deciphering the neuroscientific mysteries of the human brain. (Castelvecchi, 2016: Oh, Scheile & Fritz, 2019).

[154] Of course, this neglects certain circumstantial concerns like a high turn-over in commercial engineering departments, or the selling of a pre-designed model from one corporation to another. In other words, it may be hard to locate the engineers who originally designed the program, but abstractly at least, some human agent is aware of its content.

Furthermore, if both transparency and accountability can be seen as instrumental values for the securing of public trust and acceptance, then the *predictability* of expert systems provides an additional incentive for their implementation. Black boxes, on the other hand, may yield unpredictable decisions, whether this is caused by the machine's own intuited decision procedure, or by the biased or incomplete data sets on which they were trained. Responsibility, in turn, is often a question of human conscience when it comes to the choice between expert systems or machine learning approaches. For instance, machine learning algorithms have become increasingly popular in the medical field, and have shown impressive results in the diagnosis and detection of diseases such as cancer[155]. Still, how does the implementation of medical black boxes affect areas of ethical salience such as meaningful patient consent, or respect for human autonomy and dignity? How is the doctor-patient relationship affected by these technologies, and does it increase human accountability in medicine[156]? While machine learning techniques have proven to be efficient, if not impressive, in areas where expert systems have failed to scratch the surface of human behavioral equivalency, it is important to recognize the inherent value trade-offs embedded in these design decisions.

In summary then, it is important to clarify two points. First, most embodied forms of artificial agents employ both expert systems and machine learning approaches in their design structure; for instance, an autonomous vehicle may employ a bottom-up approach to the identification and classification of its environment, but may nevertheless employ an expert system for its tactical (and perhaps ethical) decision-making. The ethical design concerns of black boxes then, are somewhat mitigated by the scope of their use in a particular artificial agent. Secondly, expert systems are not without their own shortcomings. There are many good reasons to employ bottom-up approaches to decision making, especially if the decision-making process is too complex to be accurately described by a top-down system. We will explore these issues in detail as they apply to the computational approaches which aim at *artificial morality*, and thus we will leave them alone for the time being.

[155] Price, 2017.
[156] John Sullins (2014) makes a similar claim in the context of robotic-assisted surgery.

4. *Conclusion*

In this first chapter, we have attempted to show the degree to which an artificial agent's design and behavior is influenced by its environment; be it in a narrow sense, through the concept of an *Umwelt*, or in a broader sense, through the notion of overarching ethical design concerns, or the AI principles that many stakeholders have proposed to orient technological innovation in a safer, ethically sensitive direction. In the first section, we compared and contrasted different definitions of an artificial agent, some which placed an accent on the agent's apparent autonomous action, and others which were quite obviously tethered to the designer (or user's) decisions and preferences. In the end, we opted for a definition that characterized the artificial agent's autonomy as an autonomy or control over the means by which the artificial agent responds to its *Umwelt*'s marks of significance. Importantly, this definition married the concept of a purpose-oriented ontology with a specific environment, and linked the attainment of an artificial agent's design objectives with its capacity to react to and learn from, act through, and interact with that environment.

In section II, we investigated the concept of an *Umwelt* further, elucidating a number of ways in which engineers can classify their environment in order to better design their machines. Most prominently, this resulted in a PEAS classification of a given artificial agent, where the classification yielded: the performance measure (how it ought to act), its agent program (the way it moves from perception to action) and its actuators and sensors (the methods by which it acts within and perceives its environment). This further specified the concept of a purpose-oriented ontology, and further underscored the importance of a (specific) environment to the successful design of an artificial agent.

Finally, in section III, we moved away from a precise typification of an agent's environment and directed our attention towards the overarching normative environment to which artificial agents likely ought to respond. While the number of so-called AI principles has increased almost exponentially in recent years, we chose to focus our attention on three popular design

principles: accountability, responsibility and autonomy. Ideally, beyond a responsiveness to its *Umwelt*, an artificial agent's purpose-oriented ontology should also meet the conditions of these three principles. We explored some of the ways in which these principles can be met or fail to be respected through four general classifications of artificial agents: embodied, virtual, deterministic and stochastic. While the respect of these principles is often best achieved at a more concrete level, we nevertheless established that embodied agents generally, in virtue of their physical presence in human social contexts, have the ability to cause physical or lethal harm to human agents. This, in turn, heightened the stringency of all three ART principles, since these machines were often tasked with the role of 'deciding how to crash' or 'deciding who to target' in military operations. Moreover, the causing of harm by an artificial agent was seen to exacerbate certain moral buffers and a general denial of (moral) responsibility on the part of users and designers in some cases; while in others, it generated a host of thorny liability questions surrounding who was to blame for so-called 'driverless accidents'. In both cases, we alluded to the idea that the current degree of respect for ART principles left something to be desired. Finally, we applied these same ART principles to the deterministic-stochastic distinction, and learned that while machine learning algorithms promise to solve certain efficiency problems inherent in expert system approaches, the black box effect of these algorithms made them a problem for overarching ethical design concerns.

In the next chapter, our goal will be to employ our definition of artificial agents in an in-depth analysis of one of its features: the capacity for autonomy. In so doing, we will begin to broach the question of artificial moral agents, or those artificial agents that are endowed with a type of artificial morality, which enables them to respond to the moral value within their environment. The notion of autonomy has often done 'definitional double-duty[157]' in the discussion surrounding artificial agents, and we will need to examine it closely, picking out its necessary parts, in order to understand the true difference between human and artificial moral agency.

[157] Bryson, 2018: Dignum, 2019.

CHAPTER

2

Autonomy & Artificial Moral Agents

The concept of autonomy occupies a unique place within the machine ethics literature. In some ways, it evokes the very utility of artificial agents: operating autonomously in their *Umwelts*, artificial agents serve to lighten the practical load of the human users who employ them. In other ways, autonomy denotes the most problematic feature of artificial agents: by acting and deliberating in ways independent from human surveillance and direct control, artificial agents may perform unpredictable or harmful actions that are ill-suited to their environments, and may cause damage to the human users who surround them. In this way, the need for the implementation of artificial morality in artificial agents is intimately tied to the concept of autonomy. To illustrate, consider the following relatively austere reflections on the subject:

Morality is a fundamentally human trait which permeates all levels of human society, from basic etiquette and normative expectations of social groups, to formalized legal principles upheld by societies. Hence, future interactive AI systems, in particular, cognitive systems or robots deployed

in human settings, will have to meet human normative expectations, for otherwise these systems risk causing harm[1].

As we expect robots to do more for us, they will necessarily become more sophisticated, operate more autonomously, and do so with open, unconstrained settings. These greater freedoms come with the increased risk of unanticipated combinations of unforeseen factors leading to hazardous situations or actually resulting in harm[2].

As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged—or will soon engage—in some form of it…because machines are becoming more sophisticated and make our lives more enjoyable, future machines will likely have increased control and autonomy to do this. More powerful machines need more powerful machine ethics[3].

It is relatively easy to identify some common themes among these avowals of the need for artificial morality. First, we can see a somewhat pragmatic espousal of moral realism: morality—whether characterized as formal legal principles, etiquette or societal normative expectations—is seen to be a very real part of human social contexts. In other words, these normative expectations—whatever their source and content—feature in the marks of significance of an artificial agent's *Umwelt*; the successful performance of an AA's task therefore hinges on its successful sensitivity, and eventual response to these features. This fits with our conception of artificial morality from the previous chapter: a process by which an artificial agent is rendered responsive to the moral value which exists within its *Umwelt*[4].

A second theme in these reflections is of a justificatory nature: the failure to respond to the moral value of an *Umwelt*, on the part of an artificial agent, generates the risk of 'hazardous situations' and harm. Then, the motivation behind the implementation of artificial morality in AAs

[1] Scheutz, 2017, 59.
[2] Winfield & Jirotka, 2017, 267.
[3] Moor, 2006, 18-20.
[4] It is likely true, however, that this moral realism defended here is of a weaker form than that which is typically espoused in philosophical literature. Indeed, the characterization of morality as 'etiquette' and 'normative expectations' remains patently non-committal about what morality ostensibly *is*: sentiments, intuitions, mores or societal conventions, or perhaps indeed an espousal of objective moral truth. An investigation of the meta-ethical assumptions at work behind the machine ethics literature will drive us away from the arguments we will attempt to propose in this chapter, however they will become pivotal in our investigation of different approaches to artificial morality in part II of this thesis.

does not derive from a will to create a perfect machine, or simply to 'optimize' its behavior. Instead, the implementation of artificial morality is justified in precautionary terms: without a responsiveness to the moral value of its *Umwelt*, an artificial agent could cause harm to human agents. We have seen some examples of this in our exploration of embodied artificial agents, when autonomous weapons must decide who to target, or when autonomous vehicles must decide how to 'crash'.

Finally, we may detect that the 'autonomy' of artificial agents can be seen to *aggravate* these precautionary concerns. As AAs become more sophisticated, are deployed more generally, or operate in evermore 'unconstrained settings', the risk of human harm is amplified, and correspondingly, the need for and justification of the implementation of artificial morality is galvanized. The more autonomous an artificial agent is, in other words, the more morally sensitive it will need to be.

In the field of machine ethics, this is a powerful and popular argument[5]. To keep track of its use within the literature, we will call it *the argument from increasing automation*, which we can more formally describe as:

The argument from increasing automation: If moral value is an inherent feature of an artificial agent's environment, then a failure to account for this moral value in the agent's decision-making will result in the risk of human harm. The more autonomous an artificial agent is, the more this risk is increased.

There is one point that we must immediately address concerning this claim. We have elected to use the term increasing 'automation' and not increasing 'autonomy' with the aim of avoiding a conceptual hang up that persistently plagues the field of machine ethics. As we shall see, there is a salient difference between autonomy understood in an anthropocentric, quasi-

[5] Indeed, in their impressive attack on the justifications of artificial moral agents in the machine ethics literature, van Wynsberghe & Robbins (2019) identify a similar trend in the literature. In alignment with their analysis, the argument from increasing automation covers three points: the *inevitability* of AMAs, the *risk of human harm* of AMAs, and the *complexity* of AMAs. The cogency of the argument from increasing automation is also thoroughly defended (by a number of prominent machine ethicists) in (Poulsen et al., 2019).

philosophical sense on the one hand, and a technocentric, engineer-oriented sense on the other[6]. By electing to use the term 'automation', we can mitigate some of this confusion. 'Automation' here refers to the *mechanization of a particular task or task environment*. In this way, a highly automated environment is one in which artificial agents are either a) ubiquitous, or b) tasked with the majority share of the decisional load inherent to an *Umwelt*. The distinction between these two vectors is important, and we can intuitively grasp this if we return to our example of the meat-rendering robot from the previous chapter.

First, recall that in our original scenario, the meat-rendering robot was only an artificial agent (and not an artificial *moral* agent), precisely because its environment was devoid of any morally salient marks of significance. In the original scenario, the robot was simply tasked with rendering meat, and sorting it into the appropriate bins. We then, however, *transformed* it into an artificial moral agent by modifying the conditions of its environment: in this second scenario, the robot was not only tasked with rendering and sorting meat, but also deciding which of the carnivorous islanders *received* which meats, and which were left to starve. Here, we can see the relationship between the first and second premise of the argument from increasing automation: a failure to respond to the moral value of an environment results in a risk of human harm, since a robot that is unable to adequately respond to the moral claim of each islander (to receive meat) will run the risk of causing human harm—by depriving some islanders of their only source of sustenance, and thus ostensibly starving them to death.

Importantly however, this undesirable and morally dubious outcome can arise through different configurations of automation: understanding this island meat factory as a *sociotechnical system*, the same risk of human harm can be achieved by a) a highly complex network of minimally intelligent artificial agents, who together fail to respond to the moral value (or 'meat claims'—if

[6] Joanna Bryson can be credited with the most explicit identification of this problem: "I believe that incoherence has been introduced to AI and robot ethics debates partly because some terms are made to do 'double duty'. For example, *conscious* and *intelligent* have fairly clear psychological and even computational meanings, but as a confound are often assumed to be core to moral obligation". (2018, 2). We posit that autonomy is another, particularly harrowing example of 'double duty', and it plays an important causal role in the debate surrounding moral agency and artificial morality design, as we will attempt to show in this chapter. To this end, Johnson & Verdicchio echo this conceptual concern: "…when non-experts hear that machines have autonomy, they attribute to machines something comparable to the autonomy humans have, something close to freedom to behave as one chooses" (2017, 11).

we can excuse the expression) of the islanders, or b) a single, highly 'autonomous' artificial moral agent, whose artificial morality fails to account for the moral value of the meat-claims of the islanders. Importantly, both possibilities prompt a precautionary ethical concern of the kind underpinned by the argument from increasing automation, but only the second prompts the need for the implementation of artificial morality as we have defined it thus far. In other words, what groups these two worlds together, as an object of ethical concern, is the *automation* of the decisional aspects which are inherent to the *Umwelt.* But what separates them is the *concentration* of this decision-making in a *specific* artificial moral agent.

There are two reasons why this distinction is important for our purposes here. First, and rather trivially, many of the arguments from increasing automation make no distinction between these two types of configuration. The starter pistol of ethical concern is fired, so to speak, as soon as a sociotechnical system fails to respond to moral value, regardless of the way in which this failure occurs. In this way, the accuracy of the design of artificial morality is drowned out by an unfocused uproar over the 'run-away' decision-making of 'autonomous' machines. There is a salient difference between the ethical design of sociotechnical systems understood generally, and the design of artificial moral agents which act in special ways *within* those sociotechnical systems. Secondly, and relatedly, this conceptual imprecision undermines the realities of the design context of many artificial agents, specifically the *incremental*[7] process of automation of which artificial *moral* agents are but one option or logical conclusion. Indeed, owing perhaps to the historical conditions of autonomous systems research—for instance, to build robots capable of distant planetary exploration— "…most autonomy research has been pursued in a technology-centric fashion, as if full machine autonomy—complete independence and self-sufficiency—were a holy grail"[8]. Thus, the notion of machine autonomy carries with it something of an *individualist* assumption: that this autonomy (and the moral concern which it causes) is concentrated in an *individual* artificial agent, and not the larger interaction of a sociotechnical system.

This individualist assumption gains support from another corner of machine ethics: the philosophical analysis of moral agency, and whether this is possible, desirable or accurately

[7] Anderson & Waxman, 2012.
[8] Bradshaw et al., 2013, 5.

attributable to artificial agents[9]. The heart of the problem hides behind a simple intuition: if artificial agents are to be autonomous actors in areas of ethical salience, then they will need to be capable of moral agency to accurately respond to the moral value of their environment[10]. In practice, however, the requirements of moral agency are often tantamount to a description of the composites of *human moral agency*, including subjectivity, consciousness, rationality, beliefs, desires, and intentionality, among others[11]. This prompts second order questions regarding not only the desirability of these machines, but whether their creation—if technological limitations permit—would be *moral* in its own right[12]. In this way, the precautionary solution to the argument from increasing automation may be the harbinger of additional moral concerns, pushing designers towards 'AI complete' artificial agents[13].

The goal of this chapter is simply to explore the territory covered by the argument from increasing automation, and to show how this territory gives rise to the concept of an artificial moral agent. We will not, in other words, provide any positive account of what should count as an AMA, or the type of agency, characteristics and capacities they should exhibit. We begin by unpacking the argument from increasing automation, looking first at the concept of autonomy from an engineering perspective, and then from a philosophical point of view. Undoubtedly, a thorough

[9] Sullins, 2006 : 2009 : Johnson & Miller, 2008 : Wendall & Wallach, 2008 : Stahl, 2004 : Scheutz, 2017 : Gips, 1995 : Nyholm, 2018 : McDermott, 2008 : Kiverstein, 2007 : Irrgang, 2006 : Himma, 2009 : Floridi & Sanders, 2004 : Drozdek, 1992 : Dennett, 1998a : Bryson, 2011 : 2018 : Bringsjord, 2008 : Beavers, 2011 : Anderson & Anderson, 2007 : 2010 : Allen, Varner & Zinser, 2000 : Abney, 2012 : Behdadi & Munthe, 2018 : Nallur, 2020.

[10] This is a somewhat constructive description of what has been called the 'standard view' in machine ethics (Behdadi & Munthe, 2018: Johnson, 1985). We will explore this view shortly.

[11] One further capacity of note is that of *akrasia*, the idea that a machine must be internally 'torn' between two competing moral claims, for instance its obligations to others and its own self-interest: "a fundamental property of ethical decisions…is that they involve a conflict between self-interest and ethics, between what one wants to do and what one ought to do. There is nothing particularly ethical about adding up utilities or weighing up pros and cons unless the decision maker feels the urge to follow the ethical course of action it arrives at" (McDermott, 2008, 95). A similar claim, focusing on regret and emotional responsiveness to moral turmoil is made by (Kuipers, 2018). We will revisit these sorts of critiques of artificial moral agents again in chapter VI.

[12] Joanna Bryson is perhaps the most renowned defender of this thesis (2010: 2018: Bryson & Kime, 2011).

[13] AI complete machines (Moor, 2006) is a term used to denote artificial general intelligence, and is something of a comical nod to the NP-complete problems of GOFAI lore. The link between AI complete technology and artificial moral agents, as we will see, has much to do with the supposed complexity of artificial morality; to simulate an authentic form of moral agency in machines, designers may have to solve the problem of general intelligence (Bostrom, 2014) or Chalmers' so-called *hard problem of consciousness*.

investigation of the concept of autonomy in philosophy has been the subject of many a thesis, and so to avoid making it ours, we will focus on what is called the 'standard view' in machine ethics, a perspective which expounds the idea of moral agency in an anthropocentric way, similar to that described in the paragraph above. After pinpointing the nexus between these two perspectives, we can then explore the various 'levels of autonomy' that have been described in the machine ethics literature; which, roughly speaking, map the degree of task sharing between humans and artificial agents in morally salient contexts.

1. *Unpacking the Argument from Increasing Automation*

We have, thus far in our analysis, consistently maintained that an artificial moral agent is a special kind of AA; one that is equipped with an *artificial morality* which allows it to respond to the moral value of its environment, but also, one that is placed in an environment of moral salience, such as the modern traffic environment, or the theatre of war. This drives the precautionary claim of the argument from increasing automation: without some type of artificial morality, artificial moral agents will fail to respond to the moral value inherent in their *Umwelt*, and in so doing, generate a risk of human harm, or more broadly perhaps 'unwanted events'[14]. While a heightened concern for moral wrongdoing is perhaps a relatively novel addition to the design process of the typical engineer, the concern to avoid unwanted events—understood in a broad sense—is not. Indeed, getting an AA to behave efficiently in real-world environments is perhaps the principal challenge of artificial intelligence, whether this manifests itself in classic computational issues such as *ceteris paribus* reasoning[15], or more physical feats like efficient locomotion in rough terrains. Importantly, the concept of *autonomy*, from an engineering perspective, tracks this practical need for efficient behavior. We will investigate two such concepts of autonomy below.

[14] Wallach & Allen identify two different ways an artificial agent can be what they call 'ethically blind': "First, the decision-making capabilities of such systems do not involve any explicit representation of moral reasoning. Second, the sensory capacities of these systems are not tuned to ethically relevant features of the world." (2012, 101). We will assume that the type of 'harm' alluded to by the argument from increasing automation covers both types of ethical blindness.

[15] Often associated with the frame problem, solving the challenge of ceteris paribus reasoning (or nonmonotonic inference) in artificial agents would allow them to avail themselves of this 'distinctively human style of mental operation', namely, the "…human talent for *ignoring* what should be ignored, while staying alert to relevant recalcitrance when it occurs" (Dennett, 1998b, 198).

1.1 *The Engineer's Concept of Machine Autonomy*

As a point of departure, we will consider Russel & Norvig's definition of autonomy, as described in the widely heralded textbook, *Artificial Intelligence: A Modern Approach*:

> To the extent that an agent relies on the prior knowledge of its designer rather than its own percepts, we can say that the agent lacks *autonomy*…A rational agent should be autonomous—it should learn what it can compensate for partial or incorrect prior knowledge…After sufficient experience of its environment, the behavior of a rational agent can become effectively *independent* of its prior knowledge[16].

Compared to other definitions of classic computational terms, Russel & Norvig's treatment of autonomy is decidedly broad[17]. In effect, we can establish only three meaningful delimitations: first, that an autonomous agent is one whose knowledge of its environment extends past that which was provided by its human programmer, second, that it is perhaps *better* that rational agents be autonomous rather than not, and third, that given sufficient environmental experience, the autonomous agent will act in a self-sufficient[18] way, mobilizing knowledge which may differ from that which was originally imparted by the human programmer. Notice too, the idea of *compensation* for incorrect knowledge imparted by the programmer. This seems to suggest a connection between the evolution of an agent's environment and the need for new (self-gained) knowledge on the part of the agent. Taken together, we might reformulate this concept of autonomy in the following way: an artificial agent is autonomous to the degree to which its knowledge a) departs from the a priori knowledge provided by the human programmer, and b) aids in the robustness of its behavior in its environment. Then, a maximally autonomous artificial agent is one whose self-gained knowledge allows a high degree of robustness to environmental change, and a minimally autonomous AA is one whose knowledge fails to yield efficient behavior across

[16] Russel & Norvig, 2013, 39.
[17] As it happens, some researchers in the field of artificial intelligence have bemoaned the lack of meaningful definitions of machine autonomy in the literature: "…machine autonomy remains an elusive and ambiguous concept even in computer science and robotics." (Noorman & Johnson, 2014, 55). See also: Froese, Virgo & Izquierdo, 2007: Ezenkwu & Starkey, 2019.
[18] Bradshaw et al., 2013.

environmental change. We can call this vision of machine autonomy *as independence*, since it is the (epistemological) independence of an artificial agent that allows it to better perform its purpose in an *Umwelt*.

Indeed, a central feature of the engineer's concept of autonomy revolves around "…the capacity to act robustly to environment variations"[19], an idea underscored by an AA's capacity for *mobility* (in the case of embodied AAs), but also their *reactivity*: the ability to perceive marks of significance, respond to changes that occur between them, and possibly to learn how best to adopt these changes. In other words, the autonomy of an AA is, in part, a measure of its robustness, or adaptability to change, and this adaptability is 'autonomous' in so far as it does not require intervention from human agents. For instance, a highly autonomous vehicle is capable of efficient action in all driving contexts and under all weather conditions, while a less autonomous AV may delegate control back to the human driver when it is unable to ascertain how best to proceed[20].

Nevertheless, this robustness need not necessarily be characterized as an independence from the knowledge imparted by the designer. In effect, there is a second view of machine autonomy, one we might call '*autonomy as provision*', wherein the human designer is tasked with predicting and adequately 'equipping' the AA with all the world knowledge, rules and processes it will need to efficiently respond to any change in its environment[21]; including far-flung possibilities, such as what to do when fish start raining down from the sky[22]. In this second sense of autonomy, the AA is capable of robust behavior when faced with environmental change, however, this behavior is still entirely determined by its a priori knowledge and programming[23].

[19] Lampe & Chatila, 2006, 4057.

[20] This idea of autonomy is perfectly encapsulated by the SAE's levels of autonomy for autonomous vehicles (SAE, 2016). Indeed, the difference between level 4 (semi-autonomous) and level 5 (fully autonomous) AVs is the latter's ability to drive efficiently in all contexts and conditions.

[21] This notion of autonomy is often prevalent in discussions concerning robotics (Nolfi & Floriano, 2000), where autonomous artificial agents are said to be 'tether free 'from human agents—all of the energy and computational requirements are 'on-board 'the robot. (Brooks, 1991).

[22] See, for instance, the *aguacero de pescado* which purportedly occurs every year in the region of Yoro, Honduras.

[23] Some authors question the so-called autonomy of AA's designed in this way: "the design…even for a simple task requires some effort and empirical trials because it is not possible to identify and specify in advance the desired actions of an autonomous agent. Additionally, the evolved agent can hardly be said to be autonomous because its behavior is largely dictated by the experimenter" (Nolfi & Floriano, 2000, 148).

Presumably, one of the motives behind Russel & Norvig's claim that artificial agents *should* be autonomous (in the sense of 'independent') had to do with the inherent complications of explicitly anticipating every possible change in an AA's environment, as the vision of 'autonomy as provision' would have it[24]. In other words, from a design perspective at least, 'autonomy as independence' is preferable to 'autonomy as provision' precisely because this alleviates the designer's pressure to accurately describe an ostensibly unpredictable and complex *Umwelt*.

Still, these two visions of machine autonomy produce very different results when evaluated within the context of ethically salient environments. Autonomy as independence, it would seem, allows the robot to act in ways which are—strictly speaking—*unpredictable* to the human designer, and in this way, allows it to respond to moral value in an independent way. On the other hand, autonomy as provision allows for a responsiveness to moral value which is very much 'tethered-to' or determined by a human programmer. This generates something of a tension between, on the one hand, the efficient design of artificial agents (endowed with autonomy as independence), and on the other, certain moral concerns which drive the argument from increasing automation. We can observe this tension in Peter Asaro's espousal of the problem:

> In the case of advanced AI, a system that learns from environmental data may act in ways that its designers have no feasible way to foresee…when AI systems are allowed to continue modifying their functions and learn after they are deployed, their behavior will become dependent on novel input data, which designers and users cannot predict or control. As a result, the behavior of the learned functions will, to various degrees, also be unpredictable[25].

In Asaro's characterization of this tension, it is unclear whether his concern is derived from the fact that such independent AI systems will be deployed specifically into environments of moral salience—a claim which tightly tracks the argument from increasing automation—or whether, more generally, the fact that complex independent machines exist at all could be seen to prompt moral concern. It is likely that our answer lies in what can count as an environment of ethical salience. We have given rather strong examples thus far—the human traffic environment, accompanied by the risk of 'driverless' accidents being one such example—but it is conceivable

[24] Froese, Virgo & Izquierdo, 2007.
[25] Asaro, 2016, 191-192.

that a great many other environments could be seen to carry some ethical salience if human harm is considered in more general terms[26]. Regardless of our substantive vision of what can count as harm, and correspondingly, what can count as an environment of moral salience, the tension between machine autonomy and moral concern is clear: efficient behavior, from an engineering perspective, pushes us towards the design of independent artificial (moral) agents capable of unpredictable behavior, but ethical concern, drives us back towards comparatively sub-optimal design structures (autonomy as provision) which aim to keep the agent's behavior tethered to the human programmer. In simpler, almost crude terms: engineers can be seen to prefer unpredictable machines, but moral concern has a vested interest in their being highly predictable, perhaps even at the cost of their overall efficiency[27]. Here, we can draw rather clear parallels with the types of ethical design concern that drove the ART principles of the previous chapter. A transparent machine, and to a certain extent an accountable one, likely expounds, or at least favors, an 'autonomy as provision' approach to efficient behavior[28].

We can further provide a slightly alarmist vision of this tension if we can accept a claim made by James Moor: "By its nature, computing technology is normative. We expect programs, when executed, to proceed towards some objective—for example, to correctly compute our income taxes or keep an airplane on course. Their intended purpose serves as a norm for their evaluation—that is, we assess how well the computer program calculates the tax or guides the airplane"[29]. Certainly, Moor intends normativity in an evaluative sense; we judge, *a posteriori*, the actions of

[26] This is perhaps a decisive point for what we consider machine ethics to be as a discipline, or perhaps even, what we consider the purpose of artificial moral agents to be. The account we have given thus far—one which was framed by the argument from increasing automation—points to a negative ethics that promotes the minimization of human harm, which minimally covers physical and lethal damages, and maximally might extend to the prevention of unwanted events, such as a thwarting of human interest or deprivation of legitimate opportunity. (Johnson, 2006, 199) for instance, gives a broad account of what can count as harm, including offensive images on a screen, a signal that turns off a life support machine, or a virus that is implanted in an individual's computer. Our analysis is not yet mature enough to tackle this question in a meaningful way, but it will become important in our analysis of the *place* of artificial morality in chapter IV.

[27] Kearns & Roth, 2019.

[28] In the second section of this thesis, we will find that there is yet another parallel tension in the design of AMAs: the rift between 'top-down' computational approaches to moral behavior—those which track autonomy as provision—and 'bottom-up' approaches, which to various degrees, accept the inherent problems of unpredictable machines as a necessary external effect of accurate (or efficient) moral behavior (Allen, Smit & Wallach, 2005 : Wallach & Allen, 2008).

[29] Moor, 2006, 18.

an artificial agent according to how we think it ought to have performed the task it was destined to undertake. However, if we reflect on the notion of a *performance measure* from the previous chapter, we can see that this claim has an internal parallel.

As we've learned, the performance measure seeks to describe, in normative terms, the ideal state of the artificial agent's environment, itself brought about by the agent's actions. For instance, for a robot vacuum cleaner, the performance measure was a 'clean floor'. 'Rationality', in turn, was seen to describe the ideal process by which an agent achieves this performance measure, the more an agent maximizes its performance measure, the more rational it becomes. Since all of these premises carry some normativity, we can summarize this argument in the following way: an artificial agent, acting rationally, does what it most ought to do to achieve its purpose[30]. Returning to our two concepts of machine autonomy, we can see in a more detailed way how this could generate moral concern. Autonomy as provision would seem to produce an agent who does what it most ought to do, according to its human programmer. However, autonomy as independence, would produce an agent who may itself decide what it most ought to do. Here again, these concerns are likely heightened in environments of clear ethical salience; for instance, it matters little whether the robot or the human programmer imparts the idea that sardines and peanut butter ought not to go together in a sandwich, but matters a great deal more when a robot must decide which starving islanders ought to receive which meats[31].

Thus, we can see that the intersection between the two concepts of machine autonomy on the one hand, and the moral concern that drives the argument from increasing automation on the other, admit different levels of moral concern. Minimally, some machine ethicists may find it problematic that any degree of 'autonomy as independence' is endowed into artificial agents whatsoever, regardless of whether they are deployed in an environment of ethical salience. This marries well with the moral concern for automation itself: that humans, over time, will lose

[30] This idea is echoed by (Hall, 2009) when he describes the *utility function* of an AI system (another term for a performance measure) as the 'ought' to which it aspires through action.

[31] Virginia Dignum likely sensed this convergence of ethical 'ought' with efficient 'ought' when she wrote: "When building artificial agents, the concern is often to ensure that the agent is effective, that is, that its actions contribute to the achievement of its goals and thus enable the advancement of its purpose. However, actions that contribute to the achievement of functional goals are not always the most ethical thing to do…That is, an effective agent is not necessarily a 'good' agent" (2019, 71-72).

increasing amounts of decisional authority over a given task environment, since these machines will make decisions based on evidence and experience which moves beyond human control and prediction[32]. Moderately, we may take moral issue with the idea that artificial agents be deployed in environments of moral *salience* whatsoever, remaining agnostic about the betterness relation between autonomy as provision, and autonomy as independence[33]. Finally and maximally, we have the claim likely held by both Peter Asaro and the brunt of machine ethicists: that it is morally problematic that machines be deployed in environments of ethical salience, and that this moral concern is heightened by the unpredictability of independently autonomous machines[34]. Discussion, resolution, and design then involves normative arguments as to a) the degree to which artificial moral agents ought to be predictable—mitigating or not certain arguments from efficiency—and b) the degree to which artificial moral agents (predictable or not) should hold decisional powers in environments of ethical salience. In other words, the ideal degree of human control over moral decision-making in a given ethically salient environment. While the argument from increasing automation likely covers all three tiers presented here, more often than not, it favors argumentation from this maximal tier.

1.2 *The Philosopher's Concept of Machine Autonomy*

The evaluation of the claims made by this maximal conception of the argument from increasing automation are vastly complicated by the addition of a problem of specifically

[32] This concern is a strict example of what John Moor has called the first 'dubious maxim' of machine decision making, namely, that computers should never make any decisions which humans want to make (Moor, 1979, 226-227). He goes on to claim that there are at least some instances in which, from a strictly moral standpoint, it would be better, all things considered, to allow machines to make decisions in the place of humans, if their competence can be proved (one of his examples concerns medical decisions). We will test this counterargument and others like it at the end of this chapter and throughout the next.

[33] This is tantamount to claiming that there are some decisions which machines simply shouldn't make (Moor, 1979). We have seen this type of claim crop up in such social movements as the 'stop killer robots' campaign, although in the machine ethics literature, it is a relatively rare and harsh stance—a notable exception being Alexei Grinbaum's stance in (Grinbaum, 2018) which advocates the removal of artificial agents from environments of ethical salience through the 'randomization' of their decisions, rendering them highly unpredictable and opaque, but simultaneously amoral to the human agents these decisions affect.

[34] Virginia Dignum underscores this point when she writes: "Note that the capability to learn, and thus to adopt its behavior, is an expected characteristic of most AI systems. By adapting, the system is then functioning as expected. This makes the clear specification of objectives and purpose even more salient, as well as the availability of tools and methods to guarantee that learning doesn't go awry" (2019, 57).

philosophical origin: the problem of establishing what can *count* as a moral agent[35]. While this line of inquiry is hardly reserved for artificial agents, it takes on a decidedly *constructive* tone in the field of machine ethics. To understand how this happens, we must first take stock of a number of basic claims. First, it is plain that artificial agents, in virtue of their status as technological (and therefore non-natural) artefacts, are, so to speak, whatever we *make* them to be[36]. In other words, compared to questions as to the potential moral status of sentient beings or other non-human entities, "…the nature of machines as artefacts means that the question of their morality is not simply a question of what moral status they deserve. Rather, at the same time that we ask what moral status we ought to assign intelligent artefacts, we must also ask what moral status we ought to build those artefacts to meet"[37]. The fact that moral status—and to this end, the types of physical and metaphysical apparatus that are required for it—remains a choice in the design of artificial agents presents a particularly novel challenge for the standard moral philosopher. It means, at least in some respects, that the enterprise of machine ethics, and the response to the argument from increasing automation, cannot be decided on *purely* critical, or analytical grounds. We are not, in other words, only fighting to discover some hidden truth about robots that may change our understanding of every-day experience; we must also decide what 'robotic truths' have enough merit to exist in the world, and how to bring them about through computational means.

Second, while ethical work in the field of machine ethics certainly takes on a refreshingly constructive tone, this does not preclude the establishment of certain standards, the brunt of which are borrowed from what has traditionally been seen to encompass the necessary and sufficient conditions for moral agency in *human beings*. Throughout philosophical history, the establishment of these standards represents an incredibly popular intersection for many deep, further questions: where morality comes from, or what it is (sentiment, emotion, intuition, rational will or reason), who can have it (autonomous agents, adults, conscious persons, the mentally impaired?), what is its structure (reasons, obligations, motivations, preferences), or even its truth value or justification in human life (morality increases cooperation, or provides a life of meaning, or doesn't exist at all), among many others. While the use of these lines of inquiry as an exercise in arm-chair

[35] Sullins, 2006: 2009.
[36] As Joanna Bryson curtly puts it: "Again, the moral status of robots and other AI systems is a choice, not a necessity" (2018, 17).
[37] Ibid, p. 3. See also Millar, 2015.

reasoning is surely enriching and meaningful, their use as design prescriptions in artificial agents proves to be complicated in the extreme. The employment of this discourse in the context of design, in other words, has an expressly practical purpose[38]: to respond to the argument from increasing automation; a constraint which some moral philosophers, as we shall see over the course of our arguments, have forgotten along the way.

Still, recent research in machine ethics has seen the emergence of what is often called the 'standard' view of moral agency[39], which regroups what we could consider to be the conventional (Western) conditions for moral agency in humans. The best elucidation of the standard view is likely found in Deborah Johnson's paper, *Computer Systems: Moral Entities But Not Moral Agents*, where she stipulates the following conditions for the moral agency of an entity[40]:

1. E causes a physical event with its body.
2. E has an internal state, I, consisting of its own desires, beliefs and other intentional states that together comprise a reason to act in a certain way.
3. The state I is the direct cause of (1).
4. The event in (1) has some effect of moral importance.

While some authors have pointed to various distinctions to cast doubt on the cogency of this account of the standard view—for instance, what can count as an event internal to an agent's

[38] Indeed, Torrance (2011), similarly differentiates between what he calls *practical* and *philosophical* machine ethics, where the latter is expressly focused on questions such as the possibility of artificial moral agency, and moral agency in humans.
[39] Scheutz, 2017: Munthe & Behdadi, 2019.
[40] We have used the abbreviated version of her argument as outlined in Munthe & Behdadi (2019, 5). The full account reads: "Contemporary action theory typically specifies that for human behavior to be considered action (and as such appropriate for moral evaluation), it must meet the following conditions. First, there is an agent with an internal state. The internal state consists of desires, beliefs and other intentional states. These are mental states, and one of these is, necessarily, an intending to act. Together, the intentional states (e.g., a belief that a certain act is possible, a desire to act, plus an intending to act) constitute a reason for acting. Second, there is an outward, embodied event—the agent does something, moves his or her body in some way. Third, the internal state is the cause of the outward event; that is, the movement of the body is rationally directed at some state of the world. Fourth, the outward behavior (the result of rational direction) has an outward effect. Fifth and finally, the effect has to be on a patient—a recipient of an action, a recipient that can be harmed or helped" (2006, 198).

body[41]—the brunt of the focus revolves around clauses (1) and (2). Specifically, a main point of contention is whether the internal states of an artificial agent could reasonably equate to the sorts of *mental states* which accompany traditional accounts of action[42]. Can an artificial agent be said to act on its intentions, and are these intentions voluntary, 'arising from the agent's freedom'? One traditional assumption here relates to an artificial agent's necessarily *syntactic* processing of information with moral import, a problem which finds its roots in Searle's famous Chinese Room argument[43]. This argument attempted to show, among other things, that an artificial agent could not be seen to exhibit any substantive *understanding* of the meaning of the information it processes, but rather, merely uses syntactic rules to manipulate symbol strings. For some authors, this is cause enough to defend that artificial agents could not be moral agents[44]. For very many others, it is the lack of consciousness[45], subjectivity or higher-order intentionality[46] that is implied by arguments like the Chinese Room that forbid artificial agents from counting as moral agents. Another branch of critics focuses, in a more Cartesian bend, on the *mechanistic* agency that artificial agents necessarily exhibit[47], which would bar them from the free action that is required for moral agency according to the standard view[48]. The arguments are as extensive as they are complex, and for the purposes of this chapter at least, we would do well not to dwell on them any further.

---

[41] Stahl, 2004 : Purves et al., 2015 : Talbot et al., 2017.

[42] Johnson, 2006: Dennett, 1998a.

[43] Searle, 1980. Searle's original account of the experiment is quite lengthy, but it can roughly be summarized as follows: Searle is locked in a room and 'given a large batch of Chinese writing'. He knows no Chinese, nor can he recognize Chinese symbols. He is given a further batch of Chinese writing—slipped under the door—accompanied with a set of rules that correlate the first batch to the second. Searle is given successive batches of Chinese writing, which he is able to 'answer' using the stipulated rules which accompany these batches. After a while, Searle gets so good at correlating Chinese Symbols that, from an external point of view, his mastery of the Chinese language is indistinguishable from that of a native speaker. Nevertheless, can it be said that Searle 'understands' Chinese? He answers in the negative, and uses the force of this argument to make convincing claims against various proponents of Strong AI (who would likely hold that Searle, or the computer program he represents, *did* indeed understand Chinese).

[44] Indeed, this is Grinbaum's (2018) central thesis. In a similar vein, McDermott affirms: "But a machine could reason and behave ethically without *knowing* it was being ethical. It might *use the word* 'ethical' to describe what it was doing, but that would be to, say, clarify lists of reasons for actions" (2008, 90).

[45] Himma, 2009: Kiverstein, 2007: Bringsjord, 2008: Sullins, 2009: Irrgang, 2006: Drozdeck, 1992: Dietrich, 2001.

[46] Dennett 1998a: 1998b.

[47] Champagne & Tonkens, 2015: Coeckelbergh, 2010: Johnson & Miller, 2008: Johnson & Powers, 2005: Sparrow, 2007.

[48] Predictably perhaps, some of the arguments proposed by the defenders of this view appear to bar humans themselves from the status of moral agent. As a case in point, John Sullins employs this very counter-argument against Bringsjord's (2008) condition of free will: "If Bringsjord is correct, then we are not moral

What imports us here instead, is the effect the 'standard view' could have on the argument from increasing automation. We can witness the impact in at least two ways. Firstly, and most obviously, if it is the case that no artificial agent could be considered a moral agent at all, then this would pose an intractable design challenge for engineers. It would mean that while ethical behavior is likely required in artificial agents (at least in environments of ethical salience), this is not possible, and thus we would need—in alignment with what we have previously called the *moderate* claim of the argument from increasing automation—to remove artificial agents from environments of ethical salience. This conclusion, whether or not it is positively defended, is often lurking in the spirit of much of the machine ethics literature. This, in turn, is primarily due to the second problem the standard view poses for the argument from increasing automation.

If the first problem exerted something akin to a *downward pressure* on AMA design, one that forbids artificial agents from becoming artificial moral agents, the second problem exerts an *upwards pressure*, which inevitably places engineers on a course towards *anthropocentric* designs of artificial moral agents. In detail, even if we can maintain that the standard view of moral agency is not attainable for current robotic technology, this does not do much to stop us from ardently attempting to *approximate* it in the artificial moral agents we build. In other words, if we can agree that the standard view provides the appropriate (albeit unattainable) conditions for moral agency in artificial agents, then we can at least simulate *some* aspects of this account in the design of AMAs. Anthropomorphic approaches to artificial moral agents constitute a pervasive soft norm in machine ethics, and vary mainly in which aspects of the standard view are seen to be either attainable or necessary in artificial agents[49].

Importantly, this upward pressure from the standard view interferes with two other approaches in machine ethics. First, what is often called the *functionalist* view, which rejects such

agents either, since our beliefs, goals and desires are not strictly autonomous, since they are the products of culture, environment, education, brain chemistry etc.,…Robots may not have it, but we may not have it either…"(2006, 154).

[49] We may find our treatment of the standard view and its repercussions to be curt or topical. These questions will consume us at great length in the second part of this thesis. For now, our goal is only to show how they impact both the emergence of artificial moral agent design, and normative questions related to the ideal degree of machine autonomy.

metaphysically thick conditions as consciousness, and often relies on levels of abstractions to argue for a reinterpretation of the concept of moral agency which would include both humans and artificial agents[50]. Second, there is the influence of empirical research into the appearance and perception of moral agency in technological artefacts[51], which, beyond providing valuable insight into human-robot interaction, also may justify a shift in thinking as to the ontological status of artificial agents[52]. The influence of this second approach is a complex one. Clearly, a willful focus on the appearance of moral agency in some artificial agents would seem to violate one of the basic assumptions of AMA design; namely, that it is no great mystery what goes on inside the minds of artificial agents, and it is up to humans to decide how best to design them, or as Joanna Bryson puts it, 'to what moral status we ought to build them to meet'. On the other hand, however, the perception of AAs as moral agents provides valuable insight into the types of *expectations* that an AA's user might have concerning its behavior, or for that matter, society at large. These expectations likely hold an intimate relationship with the types of moral behavior an AMA ought to exhibit, or at the very least, indicate certain acceptability norms that engineers may wish to honor in AMA design[53].

To summarize then, the philosophical interest in the moral status of machines has multiple, if not mutually incompatible effects on the question of machine autonomy, and the argument from increasing automation. Firstly, it may exert a *downward pressure* on artificial agent design, denying the very possibility of artificial moral agents, and thus bolsters the moderate view of the argument from increasing automation: that artificial agents should not be deployed in environments of moral salience. This may prove to be sub-optimal, since it negates the impetus of

[50] This approach is championed by Floridi & Sanders (2001: 2004), and it is they who coined the term 'mindless morality' as the type of agency potentially attributable to computational artefacts (2004, 351). For other functionalist approaches see (Sullins, 2006: Brooks, 2003: Prescott, 2017).

[51] Malle et al., 2016.

[52] Approaches like these tend to focus on the phenomenological and relational aspects of human-robot interaction, as a way to escape the unsatisfactory binary choice between 'moral agent' and 'simple tool'. The goal being, as Mark Coeckelbergh maintains, to shift the focus towards "…empirically informed anthropocentric ethics that aims at understanding and evaluating what robots do to humans as social and emotional beings in virtue of their appearance…" (2009, 219).

[53] An interesting example of the combination of a functional and empirical approach to the attribution of moral status to AMAs is found in Allen, Varner & Zinser 's *Moral Turing Test*. This approach is proposed to "…bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human 'interrogators cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent" (2000, 254).

machine autonomy, and does nothing to minimize human harm. Secondly, the question of moral status exerts an *upward pressure* towards the anthropomorphic design of AMAs, advocating for the necessity of metaphysically thick dispositions such as subjectivity, free will and consciousness, which may according to some, pose supplementary ethical problems[54]. Further, these concerns are quelled neither by functionalist accounts of moral agency, which while according moral status to artificial agents, do little to direct their moral design[55]; nor empirical research as to the *appearance* or attribution of moral agency in robots, which may track acceptability more than moral value.

Despite these complications however, the standard view and the general philosophical discourse which it underpins, has made human moral agency the *paramount point of comparison* across all possible forms of artificial moral agents. As we will see, the standard view sits at a critical position along the continuum of machine autonomy, flanking on one side, ethically competent agents which are nevertheless devoid of moral responsibility, and on the other side, super intelligent machines, potentially capable of moral knowledge which extends past that of humans. While 'levels of autonomy' have often felt the sting of reductionist objections in the literature[56], it is nevertheless plain that many authors, *nolens volens*, advocate an incremental or scalar approach to artificial moral agents.

2. *The Emergence of Artificial Moral Agents: The Machine Autonomy Continuum*

[54] It is a widely held view, for instance, that it would be wrong to construct machines which had a capacity for suffering (Frey, 2008 : Bryson, 2010 : 2018), especially if these machines were designed for deployment in tasks surrounding the 3 D's of robotics (dull, dirty and dangerous jobs). Additionally, there are legitimate concerns as to the potential patiency of artificial agents endowed with stronger forms of the standard view's account of moral agency (Gunkel, 2012), as well as their corresponding legal status (Levy, 2009a: 2009b). Nevertheless, anything other than a cursory discussion of these issues, while fascinating, would carry us too far from our objectives here.

[55] This point is likely most pertinently expressed through the ongoing debate surrounding the cogency of a 'moral Turing test' (MTT) (Allen & Wallach, 2008: Allen et al., 2000: Stahl, 2004: Gerdes & Øhrstrøm, 2013). Since this test is concerned, in ways analogous to Turing's original test (Turing, 2004/1950), with the detection of a capacity via its successful imitation rather than its internal validation (Arnold & Scheutz, 2016), it can at best provide only shaky reassurance that the AMA in question effectively acts 'as moral as' the human player, without providing much explanation as to why this is so.

[56] Bradshaw et al., 2013.

In the modern world, there are many types of technological artefacts which could conceivably entertain a relationship with moral value. In effect, if we view technological artefacts as the product, or perhaps the instantiation of various human intentions[57], it should come as no surprise that those intentions aim at specific goals and purposes, which may privilege certain values over others, or benefit some over others. Undoubtedly, this happens at a global level: smartphones and communication technology allow unprecedented levels of coordinated action and awareness, but this very same technology threatens to carve a 'digital divide' across individuals of different generations, socioeconomic statuses, perhaps even across nation states. It appears then that all the world's a stage for morally salient environments, and all the men and women merely players. Clearly, however, the 'world' is too broad a context for any meaningful analysis to take place.

The argument from increasing automation, in turn, provides us with a more restrictive vision of what could count as a morally salient environment. Specifically, it provides two limiting criteria: first, that moral value should be an inherent feature of the agent's environment, and secondly, that a failure to account for this moral value will lead to a risk of human harm. This aligns with a popular definition of artificial moral agents, which defines them as "…the class of entities that can be involved in moral situations, for they can be conceived as…moral agents (not necessarily exhibiting free will, mental states or responsibility, but as entities that can perform actions…for good or evil)…"[58]. This is a decidedly functionalist definition of artificial moral agents, since it quite clearly remains agnostic as to the degree to which artificial moral agents do (or ought to) approximate the standard view of moral agency. We can say that this definition—and the analytical space it provides—covers the brunt of the various levels of machine autonomy that are discussed in machine ethics literature. There are six levels, all of which describe some relationship between a technological artefact and moral value, and all of which presuppose the deployment of an AMA in an environment of moral salience. Figure 1 illustrates these levels, and provides some common terms, examples and distinctions used to discriminate between them.

[57] Johnson, 2006: Johnson & Verdicchio, 2017.
[58] Siciliano & Khatib, 2016, 12.

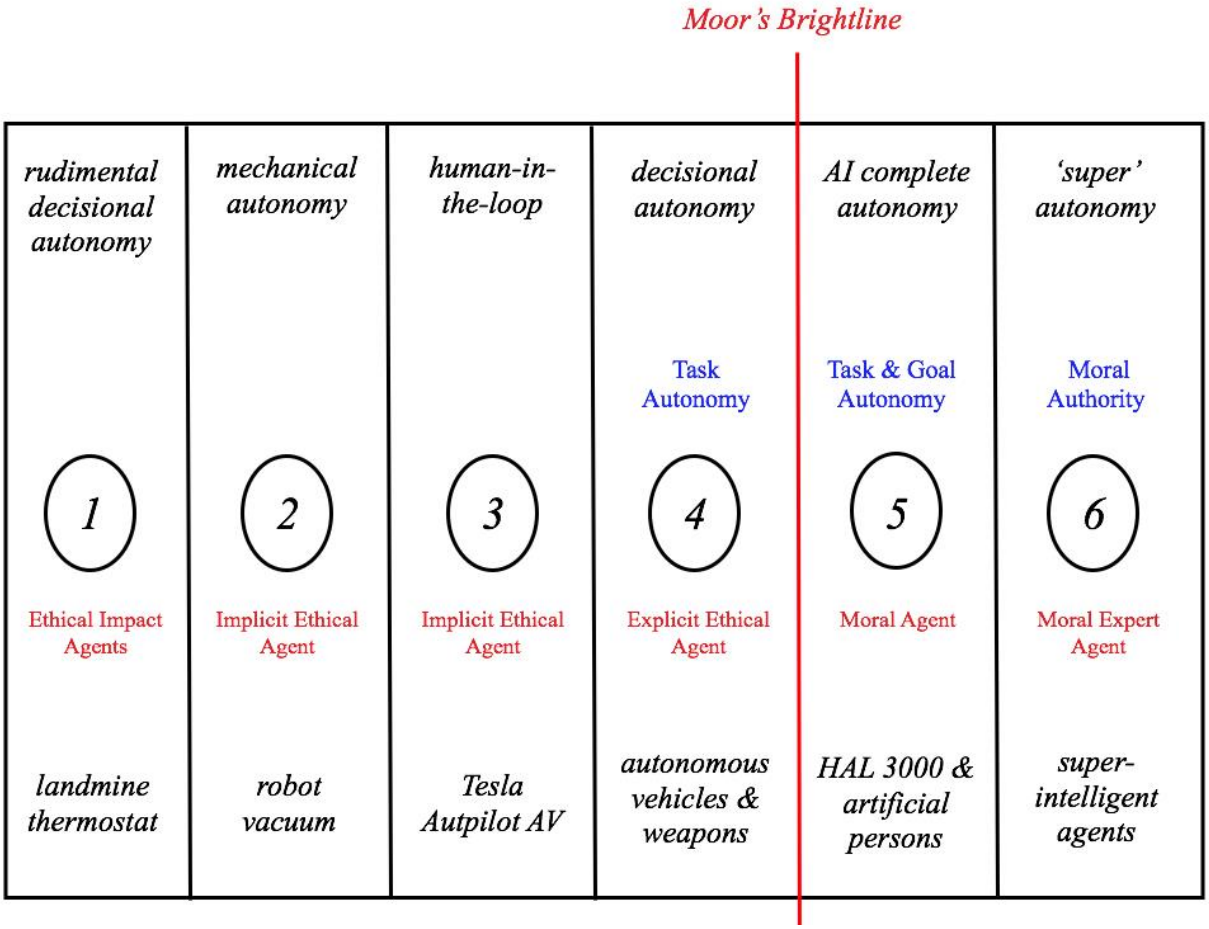


Fig. 1 - the scale of machine autonomy

Before we can address these levels and their various conditions, it is important to ask one preliminary question: what does this scale track? In an intuitive sense, the scale clearly tracks various levels of automation of a task environment. Conceiving of this task environment as a sociotechnical system in which human and artificial agents act, levels such as 1 and 2 leave significant room for human agency. The artificial moral agent, in other words, does not serve a prominent role in many decision-making tasks, but can still be seen to have an ethical impact on the moral value of the environment. At higher levels (4-6), it is clear that this ethical impact takes on a decisional nature; the AMA may not only have a moral impact susceptible to cause human

harm, but its decisions themselves may be of moral import. We have previously spoken of a *responsiveness* to the moral value of an *Umwelt* to denote this decisional capacity[59].

In another sense, this scale can be seen to track the approximation of the standard view of moral agency. Beginning in lower levels, the agent's moral capacity increases from impact to responsiveness, and as we shall see, this is typically accompanied by the need for metaphysically thick dispositions such as consciousness and moral responsibility. Level 5, in turn, exemplifies the conditions of the standard view, and is a rank which is—for the time being at least—reserved for fictitious artificial persons, characters from science fiction, and human moral agents. Additionally, beyond mental states or metaphysically thick dispositions, this autonomy scale can likewise be seen to track the degree of human moral responsibility for AMA actions—an aspect which is intimately tied to the standard view[60]. At lower levels, moral responsibility for the human harm caused in an environment of ethical salience is easily attributed to human agents, be they users, designers or manufacturers. At levels 5 and 6 however, the standard view leads us to believe that these agents are (or could be) the principal locus of moral responsibility for their actions. The dividing line between human and artificial moral responsibility often sits between levels 4 and 5[61] in the literature, and is represented by Moor's bright line in figure 1.

Finally, the scale can be seen to track the engineer's concept of machine autonomy, albeit in less straightforward ways. Levels 5 and 6, for instance, clearly presuppose an 'autonomy as independence' view of artificial moral agents. This we can deduce from the standard view's condition of free will, or freedom of action. If an AMA at this level were tethered to the a priori knowledge of its designer—as 'autonomy as provision' would have it—then it could not be said to be capable of anything beyond mechanistic agency, and thus, it would fail to meet the conditions

[59] Danaher captures this track of the autonomy scale when he writes: "Autonomy is a gradient concept which denotes the amount and level of actions, interactions and decisions an agent is capable of performing on its own" (2016, 101).
[60] Sullins, 2006: Veruggio, 2005: Moor, 2006. Sven Nyholm summarizes this point nicely when he writes: "In order for it to make sense to think that there might potentially be a responsibility-gap here, it would seemingly need to be the case that the unpredictability and lack of control depend on the presence of a significant form of autonomy or agency in the technology. Therefore, in order for a robot or automated system to pose a challenge to human responsibility, it needs to be an autonomous agent in some non-trivial sense" (2018, 1208).
[61] Peter Asaro, for instance, calls this passage the 'critical threshold of moral responsibility' (2008).

of the standard view. Conversely, levels 1 and 2 likely presuppose an 'autonomy as provision' approach to AMA design, since these levels often seek to describe what we have called *minimal* artificial agents whose design architecture is comparatively simplistic. Levels 3 and 4, however, pose something of an open question, since it is precisely among these levels that the decisional capacity of an AMA begins to take on ethical importance. Asaro's concern for unpredictable machines, and the maximal claim of the argument from increasing automation, then apply expressly to these two levels, and in a more prospective sense, to levels 5 and 6. Among these levels, in other words, the question of what a machine 'ought most to do' is a definite object of moral concern, which may be exacerbated by the machine's capacity for autonomy as independence. With these bearings in mind, we can now address each level in more detail.

2.1 *Levels 1 & 2: Rudimental Decisional Autonomy & Mechanical Autonomy*

While we don't often conjure images of thermostats and landmines when reflecting on the moral impact of emergent technology, some authors have sought to include these types of minimal artificial agents in their analysis of artificial moral agency. The principal reason for this is likely argumentative—so as to show 'where machine autonomy begins'—but this inclusion can also serve to underscore the role of human responsibility in a sociotechnical system. Indeed, landmines clearly meet the minimal requirements of artificial moral agents: they are technological artefacts that can be seen to have a salient moral impact on their environment (by causing human harm), and their environment of deployment itself is subject to ethical salience. In this spirit, James Moor has attributed the title of 'ethical impact agents' to these more simplistic machines[62]. For Moor, this ethical impact is rooted in one of the principal purposes of technology itself: to replace human agency in jobs that are dull, dirty or dangerous. To this, he adds the additional category of *demeaning* work, and uses the example of robotic camel jockeys to illustrate how these minimal artificial agents can have a hand in moral progress[63].

[62] Moor, 2006.
[63] In detail, these robotic camel jockeys replaced the human jockeys who traditionally raced camels. The condition of these young camel jockeys was seen to be akin to slavery, and so through the deployment of artificial agents, moral progress has been made.

In a similar vein, Wendall & Allen elucidate the concept of *operational morality* to describe minimal artificial agents in ethically salient contexts. This type of morality is reserved for machines with both limited *autonomy* (in the sense of either independence or provision), and importantly, a lack of *sensitivity* to the moral features of their environment. They give the example of a digital breathalyzer installed in a vehicle which tests the blood alcohol content of its user: "…a breathalyzer equipped car might prevent you from starting it, but it cannot tell whether you are bleeding to death in the process. Nor can it appreciate the moral significance of its refusal to start the engine"[64]. Here again, the notion of ethical impact—and not ethical responsiveness or sensitivity—is clearly a definitional feature of this level of autonomy.

Nevertheless, some relatively complex level 2 AMAs may fall into a secondary categorization which Moor makes, a level which confers the status of *implicit ethical agent.* The difference between ethical impact and implicit ethical agents is one of design. As Moor sees it, "…computers are implicit ethical agents when the machine's construction addresses safety or critical liability concerns"[65], a task which is accomplished by the creation of "…software that implicitly supports ethical behavior, rather than by writing code containing explicit ethical maxims. The machine acts ethically because its internal functions implicitly promote ethical behavior—or at least avoid unethical behavior"[66]. Moor gives the example of an automated bank teller to typify the category of implicit ethical agents, since the disbursement of money to different clients could be seen to carry ethical salience. The machine should, for instance, be designed to securely and accurately identify each prospective user through a passcode or PIN identification, and should perhaps be equipped with cameras so as to ensure that a withdrawal is not being made under duress.

Upon arriving at the level of implicit ethical agents then, we can begin to see the precautionary force of the argument from increasing automation. The key lies in the correlation between function, environment and autonomy. If a machine is endowed with a sufficient level of autonomy (understood as either provision or independence), and its purpose within an environment

[64] Wallach & Allen, 2009, 101.
[65] Moor, 2006, 19.
[66] Ibid., p. 19.

opens it up to the possibility of generating human harm, then the design of the artificial agent itself must preempt these ethical concerns, precluding it from causing any 'unwanted events'. We can view this as a further entry point of what we have previously called *ethical design concerns*. These concerns set out to limit the design structure of an artificial agent so as to render it conducive to ethical standards and principles, such as transparency or accountability. It does not however, specifically mandate the mobilization of moral content in the AA's agent program or software[67]. This is in stark contrast to the moral arguments we could likely level for or against ethical impact agents. Indeed, the ethical design of these types of agents has more to do with *deployment* than programming or design. It is likely immoral, in other words, to place landmines around public parks; but this immorality can hardly be mitigated by the ethical design of landmines: the simplicity of these machines precludes both ethical design limitations and any explicitly ethical programming. Instead, it is the very implementation of these technologies in inappropriate environments that is the causal force of ethical concern, and not the minimal autonomy they could exhibit in 'deciding to explode'[68].

To keep track of these ideas, we might distinguish these two types of AMAs in the following way: ethical impact agents (level 1) are subject to what we could call a morality of *use* or implementation, where it is up to human authority to decide where and when these agents could cause human harm or benefit in an environment of ethical salience. Level 2 or implicit ethical agents, on the other hand, are subject to what we could call a morality of *design*, which denotes the programmer's duty to ensure that the design structure of the artificial agent meets appropriate moral principles and standards, thus avoiding any risk of human harm. Importantly, while these two levels may differ in moral complexity, the moral responsibility for the use and design of both these types of agents, as well as any harm they could bring about, rests with human agents. As Wallach & Allen see it, "…their moral significance is entirely in the hands of designers and users"[69].

[67] "Machines' capability to be implicit ethical agents doesn't demonstrate their ability to be full-fledged ethical agents. Nevertheless, it illustrates an important sense of machine ethics. Indeed, some would argue that software engineers must routinely consider machine ethics in at least this implicit sense during software development" (Moor, 2006, 19).
[68] Asaro, 2008.
[69] Wallach & Allen, 2009, 104.

2.2 *Level 3 & 4: Human-in-the-loop Technology & Decisional Autonomy*

Level 3 artificial moral agents occupy an interesting place in the autonomy scale, arguably because the scope of human-in-the-loop technology is expansive, and has deep roots in the history of robotics. In a historical sense, human-in-the-loop technology is equivalent to telerobotics: "remotely controlled machines that make only minimal autonomous decisions…they do not need complex artificial intelligence to run, its operator provides the intelligence for the machine"[70]. One infamous example is NASA's Mars Rover, and another more current example is telerobotic surgery. In the past, the inherent limitations of telerobotic technology like the Mars Rover provided the principal driving force of the engineer's impetus towards highly autonomous machines[71]: if distances were too great between the robot and the operator, or if communication was difficult, then the robot may cease to be an efficient agent *simpliciter,* since, severed as it is from its human operator, it would be unable to move autonomously through its environment. If the artificial agent suffers from such a severe lack of autonomy, then it is likely appropriate to consider it an ethical impact agent, falling under the heading of *morality of use* in our ethical appraisal[72]. However, phenomenon such as robot-assisted surgery, or remotely operated military drones constitute far more complex examples of human-in-the-loop technology, and as such, require a more careful analysis.

The hallmark of human-in-the-loop technology, from a conceptual standpoint at least, lies in the *hermeneutic* relationship it generates between its user and the world. In other words, these technologies play the role of an in-between, or mediator that allows human agents to interpret the world without having direct sensory access to it[73]. In modern times, the human perception of these worlds is typically contingent upon the use of these technologies, which is to say that they provide an otherwise impossible vision of a specific environment; nanotechnology and distant planetary

[70] Sullins, 2006, 25.
[71] Russel & Norvig, 2013.
[72] Indeed if anything, morality of use is better suited to telerobotics than landmines, since the human decision-maker has more minute control over the precise agency of the AMA.
[73] Coeckelbergh, 2011.

exploration are two such examples of this type of relation. In this way, these artificial agents serve an assistive[74] role in at least one of two ways: in less sophisticated instances, they allow the extension of human agency into hitherto inaccessible environments, and in more sophisticated instances, they may provide various suggestions, improvements or decision-assistance to their operators. Two examples, respectively, are the robotic surgery assistant Da Vinci, which has been in use in American hospitals since the year 2000[75], and the capsule robot at Boston Children's hospital[76]. The shift from primitive to sophisticated human-in-the-loop technologies is obviously one of autonomy: the more the robot is capable of adaptability or robustness to environmental change, the more autonomy it has over the task environment, and the less essential the human operator becomes. In the case of medical robots, this shift manifests itself in the type of surgery performed. Comparing CT scanners to systems like Da Vinci, John Sullins writes: "…the CT scanner has little direct contact with the patient whereas the da Vinci is very active through the surgical process. Thus we see that when an action is safe and routine, it can be more readily automated but if the surgical action is risky and requires a lot of cognitive skill to perform, the machine must be far less automated"[77].

Interestingly, while the human operator may not be essential from a technical standpoint, it may be the case that he remains essential from an *ethical* standpoint. As Sullins' quote suggests, if the moral risk of decision-making and autonomous action is high, this may place design constraints on the machine that fail to exhaust its potential for autonomy. In this way, the nature of the task, its inherent moral risks, and the limits of technology all figure into the choice of power-sharing between operator and robot. To this end, a similar parallel exists in the field of autonomous

[74] Dignum, 2019.

[75] According to its makers, "…The da Vinci Surgical System is a tool that utilizes advanced, robotic technologies to assist your surgeon with your operation…It does not act on its own and its movements are controlled by your surgeon…The da Vinci Surgical System has…special instruments and computer software that allow your surgeon to operate with enhanced vision, precision, dexterity and control…And, da Vinci's software can minimize the effects of a surgeon's hand tremors on instrument movements". The da Vinci ® Surgery Experience, Intuitive Robotics fact sheet retrieved 3/4/20 at: https://www.steliz.org/www-seb/media/SEB-PDF-Documents/da_Vinci_Surgery_Facts_Sheet_191548(1).pdf

[76] "…a faceless white robotic cylinder about the size of a breath mint, attached to the end of a catheter" (Svoboda, 2019), this tiny robot promises to autonomously repair leaking heart valves, guided only by its vision and touch sensors. While it is still in clinical trials, the robot has the potential to become, as one engineer puts it, "the cruise control of surgery".

[77] Sullins, 2014, 5.

weapons, where lethal decision-making, while ostensibly executable by autonomous drones, is still necessarily confirmed or validated by human agents in the military chain of command[78]. Many authors have pointed to the sub-optimality these design choices create in the power-sharing between humans and robots. From one side, the desire on the part of humans to remain 'in the loop' may render the entire enterprise unsuccessful, dangerous, or less efficient[79], and from the other, the autonomous capacities of advanced AAs may push humans out of the loop entirely, generating novel moral quagmires[80].

In other words, there are at least three types of artificial moral agent that can find themselves in the third level of the autonomy scale: a) ethical impact agents with minimal autonomy and no sensitivity to moral value, b) implicit ethical agents with greater autonomy, and an implicit sensitivity to moral value or overarching ethical design concerns, and c) a highly autonomous agent capable of efficient ethical decision-making, but whose decision-making is nevertheless truncated by human authority, with variably ethically troubling results. The type of AMA depicted in (c) marks the passage from level 3 to level 4 on the autonomy scale, denoting the shift from implicit ethical agent to what is called an *explicit* ethical agent in the literature. In effect, explicit ethical agents can exist at both levels of autonomy, and the passage from one to the next is conceptually distinguished by the AMA's capacity for *task autonomy*. We will treat each of these concepts in turn.

[78] Asaro, 2008. See for instance, the American government's design safety precept (DSP) 15: "The firing of weapons systems shall require a minimum of two independent and unique validated messages in the proper sequence from authorized entity(ies), each of which shall be generated as a consequence of separate authorized entity action. Both messages should not originate within the UMS launching platform". (Arkin, 2009, 27). An *UMS*, or unmanned system, is a synonym for artificial agent in our case.

[79] A good example of this can be found in autopilot autonomous vehicles (typically called 'level 3' AVs) (SAE, 2016). In this scenario, the driver is prompted to take back control of the vehicle when accident scenarios are likely or imminent, which typically results in a failure on the part of the human driver to quickly and efficiently react to the situation at hand.

[80] This is the pessimistic prognosis of most future military sociotechnical systems: "the military systems (including weapons) now on the horizon will be too fast, too small, too numerous, and will create an environment too complex for humans to direct. Furthermore, the proliferation of information-based systems will produce a data overload that will make it difficult or impossible for humans to directly intervene in decision-making" (Adams, 2001, 2). Some authors fear that this will lead to a decrease in moral accountability—the effect of moral buffers (Cummings, 2006, 35), while others fear these systems will incentivize proliferated warfare, by lowering the barriers of entry into conflict (Sparrow, 2007).

James Moor coined the term 'explicit ethical agent', a concept which is now widely used in the machine ethics literature. His original definition, unfortunately, lacks some useful clarity. In effect, Moor does not provide a thorough account of the necessary and sufficient conditions of explicit ethical agents, and tends instead to beg the question:

> Can a machine represent ethical categories and perform analysis in the sense that a computer can represent and analyze inventory or tax information?…Can a machine represent ethics explicitly and then operate effectively on the basis of this knowledge?…What would such an agent be like? Presumably, it would be able to make plausible ethical judgements and justify them. An explicit ethical agent that was autonomous in that it could handle real-life situations involving an unpredictable sequence of events would be most impressive[81].

A successful filtering of his speculative tone reveals a number of potential conditions: a) an explicit representation of moral value in the agent's program, b) a capacity for plausible (or predictable) ethical decisions, and c) a capacity to justify these decisions. We might add d) a high degree of autonomy, understood either in terms of provision or independence, even if Moor seems to regard this as supererogatory. Importantly, Moor maintains that explicit ethical agents are not equivalent with human moral agents, nor do they necessarily meet the conditions of the standard view[82]. What we are left with, at an abstract level, is a capacity for *moral responsiveness*, an idea which is echoed by equivalent terms in the literature. Wallach & Allen for instance, equate explicit ethical agents with the concept of *functional morality*, "…where the machines have the capacity for assessing and responding to moral challenges"[83]. Bello & Bridewell, in an illuminating article, describe their 'agents of a third type' in a way which likely coincides with explicit ethical agents: "Agents of the third type are capable of committing to choices with respect to norms…[their] explicit encoding makes the norms available for reasoning, comparison, and interchange based on the dynamics of the situation"[84].

---

[81] Moor, 2006, 20.

[82] "We won't resolve the question of whether machines can become full ethical agents by philosophical argument or empirical research in the near future. We should therefore focus on developing limited explicit ethical agents. Although they would fall short of being full ethical agents, they could help prevent unethical outcomes" (Ibid., p 20).

[83] Wallach & Allen, 2008, 106.

[84] Bello & Bridewell, 2017, 28. They go on to describe some supplementary behaviors such as a) violating cost minimization in light of circumstances in which the expected costs of performing various actions are unknown whereas the expected utilities of potential outcomes are well characterized, b) decision-making

What we can see in these definitions is the emergence of a responsiveness to the moral value of an artificial agent's *Umwelt*, which we can characterize, internally, as a *recognition* of certain moral facts and features present in the agent's environment. Importantly, this does not imply a deep *understanding* of moral features and facts as such by the agent[85], but rather, it implies that the agent program (or software) implemented into the AMA allows it to move from perception to action, precisely by recognizing certain percepts as being of *moral salience*, and by addressing these percepts in a specific way. This is the basic structure of what we have called *artificial morality*: a decision procedure which allows an AMA to *respond* to the moral value of its environment, by perceiving and *recognizing* this value as such, and by applying a *method* which allows it to respond to this value in a normative way. Put very bluntly, artificial morality describes both *what* is of value in the *Umwelt* of an artificial moral agent, and *how* (or how much) it ought to value these aspects of its environment[86]. Thus, for our purposes here, we can say that an explicit ethical agent is one which is equipped with artificial morality.

This, at least, appears to satisfy condition (a) of Moor's definition. Condition (b), the capacity for plausible or predictable ethical decisions, has much to do with the *type* of artificial morality which is implemented into the AMA, a subject which we will broach in the second section of this thesis. Here, we will simply say that there are some approaches which yield more predictable decisions than others, just as there are some approaches which enable more flexible and robust decision-making than others. The satisfaction of (b) then, has much to do with the choice and accurate design of artificial morality. Finally, condition (c), a capacity for justification, would seem to align with the requirements of ethical design concerns such as accountability and transparency. At the very least then, an explicit ethical agent should be able to explain its decisions in ways that are transparent and intelligible to relevant human agents such as designers and users. This again, we will contend, is achievable if the designer can state explicitly the nature of the

in situations where the utilities of two different actions are incommensurate, or c) a commitment to the pursuit of their goals in the face of distractions and situational reappraisals.

[85] In other words, the AMA's processing of information remains syntactic, in alignment with the Chinese Room critique supplied in the first section of this chapter.

[86] Our explanation of artificial morality must remain cursory here, we will address it directly in the next chapter, and explore various proposals in the literature in part II of this thesis.

artificial morality that is implemented into the AMA. Beyond this, the precise nature of the justification is only practically discernible when additional information is known: the type of relationship the AMA holds with its users, and the type of agent it is (embodied, virtual, autonomous vehicle, etc.) being two prominent criteria.

Condition (d) a high degree of autonomy, in turn, requires some supplementary attention. Clearly, at least a part of what Moor intended aligns with the engineer's concept of machine autonomy as we have defined it in this chapter. To this end, it would seem that level 4 explicit ethical agents likely possess an autonomy as independence, if they are to be successful actors in dynamic and complex environments. This, in any case, is what Moor deemed to be the 'most impressive' design scenario[87]. We should however, by way of cursory clarification, maintain that while the autonomy (as independence) of explicit ethical agents is likely very high, this does not imply that these agents *by themselves* discover what they most ought to do. As Bello & Bridewell claim, explicit ethical agents are capable of normatively sensitive action via explicit coding in their programming, allowing them to consider, compare and reason about the moral value in their environments, and changes within it. Importantly, these norms are not discovered by the robot itself, they are instead implemented by the human programmer in a traditional, 'top-down' fashion, as costs, constraints, or reward functions in the AMAs cognitive architecture[88]. This 'top-down' or symbolic AI assumption is present in all of the definitions of level 4 explicit ethical agents we have explored thus far, but there are other approaches in which the robot itself may decide, to a greater extent, what it most ought to do. In this way, the autonomy of a robot *qua* response to moral value is somewhat separate from its autonomy as independence: a highly independent AMA can possess a human-given artificial morality that strictly constrains its response to moral value, or it

[87] John Sullins, in speaking of level 4 AMAs, likewise hints at an increasing degree of autonomy as independence in his elaboration of the concept of 'effective autonomy': "I mean to use the term 'autonomy' in an engineering sense, simply that the machine is not under the direct control of any other agent or user…If the robot does have this level of autonomy, then the robot has practical independent agency. If this autonomous action is effective in achieving the goals and tasks of the robot, then we can say the robot has effective autonomy. The more effective autonomy the machine has, meaning the more adept it is in achieving its goals and tasks, then the more agency we can ascribe to it" (2006, 28).

[88] we are clearly alluding to the type of artificial morality implemented into the AMA, of the top-down, bottom-up or hybrid variety as explained in (Allen, Smit & Wallach, 2005). We will spend much time with these concepts in chapter IV.

can have a more bottom-up or self-directed response to moral value, following for instance, an embodied or machine learning approach.

With this in mind, we can now clarify some further differences between level 3 and 4 artificial moral agents. We have maintained that it is possible for a level 3 AMA to nevertheless be an explicit ethical agent, a highly independent agent who decides, typically through explicit normative programming, what it most ought to do. The role of the human operator in this case is somewhat symbolic or superfluous: he is often present so as to verify or supervise the autonomous decisions of the robot. What this type of level 3 agent lacks, and what a level 4 explicit ethical agent necessarily has then, is the capacity for *task autonomy*. A level 3 AMA is not autonomous in the pursuit of its tasks, while a level 4 AMA necessarily is. Task autonomy, in turn, denotes "…the ability of a system to adjust its behavior, by forming new plans to fulfil a goal, or by choosing between goals"[89]. We can see that this concept overlaps with aspects of both autonomy as independence and artificial morality, however its application to our analysis is perhaps not very clear. Virginia Dignum, the author to which this concept is owed, further qualifies that task autonomy is "…relative to the means or instrumental sub-goals accessible to the agent"[90], and uses the example of autonomous vehicles to clarify her argument. In autonomous vehicles, the passenger provides the destination which he would like to reach, and the AV computes the most efficient route by which to reach it. Task autonomy, then, relates to the AV's capacity to plan the most efficient route to the destination, a route which may change or develop in light of the dynamics of the vehicle's environment, but which may also be affected by the AV's artificial morality: it may forbid the AV from planning routes through dangerous portions of a city, it may limit the maximal speed at which the AV can proceed, and it may require the vehicle to stop for j-walking pedestrians even if the passenger would prefer otherwise. Thus, task autonomy simply denotes the AMA's ability to decide for itself the most efficient route to achieve its goal, where the goal itself is provided by the designer or user. Artificial morality, in turn, may put additional constraints on the means by which it achieves this goal.

[89] Dignum, 2019, 21.
[90] Ibid, p. 21.

In light of this, we may initially confuse level 3 telerobots with level 4 agents with task autonomy, after all, both seem to be constrained by a human agent. However, such an assumption would be erroneous, since a level 3 AV would require not just positing the goal of the system (to reach a certain destination), but also would require that the passenger manually *drive the vehicle* for at least some portions of the route, taking over for instance, in critical situations or sub-optimal weather conditions. Thus, the notion of autonomy is further splintered by the degree to which a human agent aids in the achievement of a system's goal. To summarize then, the move from level 3 to level 4 on the autonomy scale covers a wide range of AMAs: first, there are level 4 explicit ethical agents, equipped with artificial morality and task autonomy, who require minimal control from human agents. Then, there are level 3 explicit ethical agents, who while likely possessing artificial morality and the *capacity* for task autonomy, are nevertheless tethered to human operators who decide not only the goal of the system, but who likely aid in the achievement of this goal in direct ways through intervention in the machine's agency. After this, there are less sophisticated level 3 implicit ethical agents or ethical impact agents, who do not possess the capacity for either artificial morality or task autonomy. If ethical impact agents and implicit ethical agents are seen to be subject to a morality of use and a morality of design respectively, we could argue that explicit ethical agents are subject to a morality of *behavior*: their decisions and actions themselves are subject to moral appraisal.

2.3 *Moor's Bright Line, Moral Agents and Superintelligent Agents*

In his seminal paper discussing the potential forms of artificial moral agents, James Moor points out a certain threshold which many in the field of machine ethics seem unwilling to cross: "Many believe a bright line exists between [implicit and explicit ethical agents] and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future"[91]. What is meant by this 'crucial ontological difference'? Moor offers two accounts. The first account holds that the

[91] Moor, 2006, 19.

bright line represents an exclusivist position on the potential moral agency of artificial agents. Since the morality of AMAs is reducible to the morality of its designers, it does not make sense to consider machine ethics as an ethics involving true moral agents. This may lead us to believe that an AMA's response to moral value in its environment is not an ethically salient situation at all, or only appears to be one[92]. The second account holds that the bright line represents a critical limit on the design of AMAs, namely that "…no machine can become a full ethical agent—that is, no machine can have consciousness, intentionality, and free will"[93]. Patently, the ontological differences alluded to by both accounts are intimately linked to the standard view of moral agency: the first describes how current AMAs fail to be the locus of their moral behavior—a condition relating to the capacity for free will in the standard view—while the second account holds that no machine could ever be the locus of its own moral behavior, and they thus necessarily fail to meet the conditions of the standard view.

A clear upshot of the bright line argument is that it reveals the basic definitional assumption of level 5 AMAs: they are 'full ethical agents' which satisfy the conditions of the standard view, and perhaps, that humans consistently meet these conditions. Importantly, the satisfaction of the conditions of the standard view does more than simply render possible the ascription of moral agency to the agent, it likewise ushers in the possibility of moral patiency, and moral responsibility[94]. It is perhaps for this reason that Peter Asaro defines this same bright line as the 'critical threshold of moral responsibility': beyond this point, the robot could be seen to be a locus for *moral*, and not simply causal responsibility for morally reprehensible events in the world[95]. From a normative perspective, many authors have pointed to the risks and shortcomings of AMA design which would aspire to this level[96], and some further conclude that regardless of the agent's level of autonomy, human agents will and must remain morally responsible for the actions of the agent[97].

---

[92] Sullins, 2006.
[93] Moor, 2006, 19.
[94] Johnson, 2006.
[95] Asaro, 2008.
[96] Bryson, 2011: 2018.
[97] A good example of this position is found in Johnson & Miller's espousal of the 'computers in society' approach to AMAs: "artificial agents will be understood to be human constructions under human control. They would always be understood to be 'tethered' to humans in the sense that they are the products of human intervention, are deployed by humans for human purposes, operate in contexts maintained by

Another way to characterize these concerns is to say that they represent firm positions on how to resolve what is often called the 'responsibility gap' in AMA action, where responsibility is typically (but not exclusively) understood to be of a moral nature[98]. Andreas Matthias is often credited with its original formulation in the context of machines, and he characterizes it in the following way:

> …certain recent developments in the way of manufacturing computerized, highly adaptive, autonomously operating devices, inevitably lead to a partial loss of the operator's control over the device. At the same time, the degree in which our society depends on the use of such devices is increasing fast, and it seems unlikely that we will be able or willing to abstain from their use in the future. Thus, we face an ever-widening *responsibility gap* which, if not addressed properly, poses a threat to both the consistency of the moral framework of society and the foundation of the liability concept in law[99].

Immediately, we should see an overlap between the argument from increasing automation and this responsibility gap. If the concern of the former was to prevent unwanted events or the risk of human harm through the implementation of artificial morality into AMAs, the latter's concern lies in preventing unaccountable or morally *unattributable* actions which risk unwanted events or human harm. Both problems—a lack of ethical agents, and a lack of morally responsible agents—are caused by the increasing autonomy and ubiquity of machines. In short then, the responsibility gap points to an empty space in the scale of machine autonomy where a clearly defined morally responsible agent ought to be, and this space minimally covers level 4 explicit ethical agents, but may, in more liberal accounts, cover the space between minimal telerobotics right up to Moor's bright line. While the discussion of moral responsibility in artificial agents has received copious attention in the literature, it is beyond the scope of this chapter to explore it in detail. For our purposes here, we must simply emphasize the correlation between the standard view, the responsibility gap, and level 4 and 5 artificial moral agents. We will attempt this now in a highly cursory way.

---

humans, and cannot function without some degree of human control (even though that control may be distant in time and space)"(2008, 128).

[98] Gunkel, 2017: Nyholm, 2018: Coeckelbergh, 2016: Dignum, 2019: Sparrow, 2007: Matthias, 2004.

[99] Matthias, 2004,178-179.

There is some evidence of philosophical and professional consensus around the locus of responsibility of levels 1-3 of the machine autonomy scale. Typically, the responsibility for the unwanted events caused by these sorts of technologies is attributed to the designer, the user, or the manufacturer. This follows what John Sullins has called the 'user—tool—victim' model, where the tool (or artificial moral agent) acts as a simple extension of the agency of the user[100]. Concern for the responsibility gap thus arises with the emergence of level 3 and 4 explicit ethical agents, who may be able to act outside of the parameters set by the human controller (or designer), albeit in a highly limited sense. Importantly, if these agents act outside of their parameters, it is not in the pursuit of their own *goals*. Instead, concerns for a missing locus of moral responsibility arise due to a) the unpredictable (malfunction) of the machine[101], or b) the mitigating factors which would appear to degrade or undermine the cogency of human moral responsibility over AMA actions—moral buffers[102] or the 'distancing'[103] problem are popular examples of this. This analytical precision, however, is lacking in the brunt of at least the *early* espousals of the problem of moral responsibility, and we can see a candid example of this with Robert Sparrow:

> …autonomy and moral responsibility go hand in hand. To say of an agent that they are autonomous is to say that their actions originate in them and reflect their ends. Furthermore, in a fully autonomous agent, these ends are ends that they have themselves, in some sense, chosen…In both of these things, they are to be contrasted with an agent whose actions are determined, either by their own nature, or by the ends of others. Where an agent acts autonomously, then, it is not possible to hold anyone else responsible for its actions…I shall argue that the more these machines are held to be autonomous the less it seems that those who program or design them, or those who order them into action, should be held responsible for their actions[104].

Notice that Sparrow's argument passes imperceptibly from the use of a philosophical concept of autonomy (freely chosen action) which is consistent with the standard view, to the use of the engineer's concept of machine autonomy, which is consistent with an independence from

[100] Sullins, 2006.
[101] Virginia Dignum supports this view in her elucidation of the state of the art of robotic moral responsibility, providing a curt response to the responsibility gap: "…basically two things can happen, either: (i) the machine acts as intended and therefore the responsibility lies with the user, as is the case with any other tool; or (ii) the machine acts in an unexpected way due to error or malfunction, in which case the developers are liable". (2019, 57).
[102] Cummings, 2013, 56.
[103] Sullins, 2006: 2013.
[104] Sparrow, 2007, 65-66.

the programmer's a priori knowledge. We could then imagine a sort of 'tipping point' at which machine autonomy miraculously translates into substantive moral autonomy, the likes of which meet the conditions of the standard view. This tipping point is likely Moor's bright line. What is important however, is that arguments like these support the idea of an *incremental approach* to moral responsibility, which labors under the assumption that machine autonomy and moral autonomy are equatable concepts. In this way, the more an AMA approximates the satisfaction of the standard view, the more morally responsible it can become. Given that artificial moral agents are still technological artefacts, and thus, are the product of human intention and construction, it seems odd to assume that metaphysically thick dispositions would simply *emerge* from an AMA's increasingly independent response to the changes in its environment. If anything, they would emerge from the *designer*'s belief that things like consciousness are necessary for full ethical agents, and his corresponding belief that full ethical agents are what is *required* for efficient action in an *Umwelt*. In response to these types of arguments, many authors have argued for a *distributed* account of moral responsibility for the morally reprehensible actions performed by highly autonomous AMAs, one which often tethers responsibility loci to the intentional actions of multiple actors in a sociotechnical system[105].

Interestingly, much of these earlier espousals of the responsibility gap find their force in the distinction between self-given and human-given ends in the agent's architecture. To this end, Virginia Dignum provides the concept of *goal autonomy* to indicate this capacity for self-determination (in a quasi-philosophical sense), and describes it as "…the ability to introduce new goals, modify existing goals, and quit active goals"[106], such as what would hypothetically occur when an autonomous vehicle informs its *passenger* where he is better off going. In a similar vein, Sven Nyholm describes this capacity as *domain specific responsible agency*[107], and provides the useful image of an AMA being able to 'stand its ground' on the basis of its own principles, and in

[105] Pettit, 2007: Johnson & Miller, 2008: Johnson & Verdicchio, 2017.
[106] Dignum, 2019, 21.
[107] "Pursuing goals in a way that is sensitive to representations of the environment and regulated by certain rules/ principles for what to do/not to do (within certain limited domains), while having the ability to understand criticism of one's agency, along with the ability to defend or alter one's actions based on one's principles or principled criticism of one's agency"(Nyholm, 2018, 1205).

the face of (human) moral criticism. Whether or not this capacity coincides with the satisfaction of the conditions of the standard view, remains for the time being, a highly speculative question.

Let us clarify our discussion up to this point. Decidedly, Moor's bright line can be seen to indicate a number of important moments for machine autonomy. First, it may indicate the tipping point at which machine autonomy can be seen to require metaphysically thick dispositions such as consciousness or free will, which may occur as a) the direct result of the evolution of the machine's autonomy as independence, or b) a positive design decision by the programmer. Second, Moor's bright line may represent a 'critical threshold of moral responsibility', beyond which an AMA is the principal locus of moral responsibility for its actions. The cogency of this critical threshold, in turn, will depend on our normative commitments to 'tethering' AMAs to human agency, our belief in the surmountability of 'Strong AI-type' problems such as consciousness, and whether we believe that a capacity for goal autonomy or domain specific responsible agency provide sufficient grounds for the attribution of moral responsibility to AMAs. Our visions of these problems will greatly contribute to our understanding of the 5th level of machine autonomy, and whether artificial moral agents could or should reach such a level. For now, conventional assumptions in machine ethics reserve this level for human moral agents only.

An interesting hiccup in this debate, which we will only begin to unpack in this chapter, is the tendency in machine ethics to espouse what we might call a *better angels of our nature* argument concerning the potential morality of AMAs. We provide a few examples below:

> If we are using human judgements to model machine judgements, then robots will inevitably incorporate the biases and inconsistencies in our own psychology: preference for people who are familiar or genetically related, ignoring the effects of our actions on people who are very distant, and relying on false beliefs about what kinds of actions are harmful.[108]

> …it is logically possible, though not probable in the near term, that robotic moral agents may be more autonomous, have clearer intentions, and a more nuanced sense of responsibility than most human agents[109].

> Some might say that only humans should make [ethical] decisions, but if (and of course this is a big assumption) computer decision making could routinely save more lives in such situations than

[108] Leben, 2018, 3.
[109] Sullins, 2006, 29.

human decision-making, we might have a good ethical basis for letting computers make the decisions[110].

…the moral environment of modern earth wrought by humans together with what current science tells us of morality, human psychology, human biology and intelligent machines morally requires us to build our own replacements and then exit stage left. This claim might seem outrageous, but in fact it is a conclusion born of good, old-fashioned rationality[111].

Organized in this highly suggestive way, the *better angels of our nature argument* seems to challenge not only the authority of human ethical decision-making, but perhaps even our very *capacity for it*. Until this point, the scale of machine autonomy, according to some, has been incrementally approaching the standard view; one long, Lamarckian scale which converges on human moral agency[112]. It would appear then, that there is no higher standard to which artificial moral agency ought to aspire. In practice however, it is true that humans often make regrettable errors in the moral choices they make, the ends they pursue, or in the way they treat others. If humans can be seen as imperfect moral vessels in this way, then this leaves open the possibility that machine morality could *improve upon human moral action*, filtering out the regrettable human errors which are seen to cause such great harm and frustration in the modern world. Importantly, the potential for human moral improvement is a question of *competence*: in the second quote for instance, the argument for the subsumption of human moral authority was contingent upon the capacities of the artificial moral agent in question. If, say, the implementation of autonomous vehicles could contribute to a significant reduction in automotive fatalities—if AVs were seen to cause less accidents or be relatively error free—then it would provide the moral grounds for a shift towards exclusively driverless cars on the world's roads[113]. In this first sense, it is not so much the *moral* capacity of AMAs that is in question, but rather their *technical capacity*: to circumvent easily avoidable and morally unfortunate behaviors that human drivers are seen to exhibit, such as texting, drinking, or holding conference calls while driving. To maintain that an AV chooses not to drink and drive because its explicit moral programming shows this act to be wrong is to

[110] Moor, 2006, 20
[111] Dietrich, 2001, 2.
[112] Joanna Bryson has criticized a similar Lamarckian scale concerning the notion of machine intelligence (2018).
[113] *A forteriori,* (Sparrow & Howard, 2017) argue that if AV competence were to exceed that of human drivers in a meaningful way, then it would be *unethical* to allow humans to drive at all.

misunderstand a great deal about morality and cars. Thus this first sense of the *better angels of our nature* argument is somewhat innocuous and relatively pervasive: technology often outperforms humans, and in this sense, it is 'better' that it does the job in our stead[114].

What would it mean, however, for machines to outperform humans *morally*? In this second sense of the better angels of our nature argument—which is alluded to in quotes 1, 2, and 4—it would seem that artificial moral agents may come to be considered as *moral experts*, relative to their imperfect human counterparts. This would open us up to a severe reversal of roles: instead of 'tethering' artificial moral agents to human normative and moral standards, we ought instead to *defer* to these machines so as to correct our own moral failings, potentially establishing new norms and principles hitherto inaccessible to humans. This of course, depends greatly on our espousal of two visions of morality: first, we must ardently espouse *moral realism*, or the claim that it is possible for every human being to be incorrect about which actions are wrong or permissible[115]. Second, we must support a specific, quite clearly *rationalist* vision of what morality ostensibly is, viewing it as something which could approximate an objective science. If we endorse these two claims, the better angels of our nature argument goes as follows:

> **The Better Angels of Our Nature Argument**: if humans are imperfect moral vessels, and artificial moral agents are immune to these human failings, then we ought to defer to their moral authority.

In the machine ethics literature, we might be surprised to find that this argument is common[116]. In this chapter, we will only briefly examine a highly hypothetical case, one which often crops up in discussions of super intelligent machines, which we have placed as the 6th degree of machine autonomy. Many of the authors of the more futurist persuasion in machine ethics have alluded to the concept of deference to machine authority. Of considerable notoriety are Nick

[114] This again aligns with Moor's first dubious maxim of machine decision-making: that computers should never make any decisions which humans want to make. He maintains that this is fallacious for precisely the same reason we have above: "If the computer's diagnosis and suggestions…would result in a significant savings of lives and reduction of suffering compared with human decision-making on the subject, then there is a powerful moral argument for letting computers decide" (Moor, 1979, 226).
[115] And, correspondingly, that it is possible for artificial moral agents to be *correct*.
[116] Leben, 2018: Gips, 1995: Dietrich, 2001: Arkin, 2009.

Bostrom and Eliezer Yudkowsky[117]. To the former, we owe the *principle of epistemic deference*: "…a future intelligence occupies an epistemically superior vantage point: its beliefs are (probably, on most topics) more likely than ours to be true. We should therefore defer to the superintelligence's opinion whenever feasible".[118]To the latter, we owe the concept of *coherent extrapolated volition*: "…our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted"[119]. Taken together, we have the makings of a real-life *ideal observer*, an impartial, omniscient, logical oracle free from attachments and commitments to any view or person. Supposing that moral realism is true, this artificial ideal observer would minimally surpass humans in its understanding of moral truth, and maximally fully understand moral truth. In both cases, human beings would have a straight-forward and strong moral reason to defer to this machine in all relevant decision-making: through the machine, they would discover what they most ought to do[120].

The proposals of these authors are as contentious as they are speculative for the brunt of machine ethicists, and we would do well to avoid rehearsing them here. Our goal is not to disprove the possibility of super intelligence, but rather, it is to expose a subtle hypocrisy in the evolution of machine ethics. At the outset, the principal motivational force of the design of AMAs flowed directly from the argument from increasing automation. The central claim of this argument was that without proper ethical constraints, artificial agents would likely cause harm to the humans their actions affected. From here, machine ethics—especially the philosophers—set out on a quest to approximate human moral agency in artificial moral agents, describing in so many ways how certain anthropomorphic features were necessary conditions of any meaningful form of moral action. Agents who did not possess these features were seen to possess, at the very most, an imperfect or superficial form of moral agency, barring them from any meaningful moral status. At

[117] We will leave Ray Kurzweil and his 'singularity' to one side for the moment.
[118] Bostrom, 2014, 5370.
[119] Yudkowsky, 2004, 53.
[120] In addition to an espousal of moral realism, this account requires what van Wynsberghe & Robbins (2019) have called stance-independent moral truths, those which do not depend on human desires, beliefs, needs etc. They argue that this account would require that we accept, "…on faith that machines are better than we are".

the very same time however, these same metaphysically thick capacities like consciousness, subjectivity, and acting on one's own motives and principles all clearly lead to occasionally sub-optimal moral decision-making in the better part of humanity. Picking up on this, some authors have seen fit to make artificial moral agents the better angels of our nature, removing frailties such as partiality and loyalty to imagined communities from the decisional architecture of these machines. Indeed, as they maintain, without this 'patch' on imperfect human behavior, robots could very well commit moral wrongs and risk unnecessary degrees of human harm. Still, in accepting this, have we not unwittingly moved from the precautionary goal of minimizing the harm generated by machines designed for specific *practical* purposes, to the goal of designing an ideal, almost *universal moral agent*, who may ostensibly know better than humans what they ought to do in any context? The most obvious instantiation of this shift is found in the dreams of futurists and super intelligent technology, but the assumptions which support it are more pervasive than we might initially imagine. As we shall see, the better angels of our nature are very much alive in the literature surrounding current robotic technology, and this may in some cases cause us to lose track of the original purposes of artificial moral agents.

3. *Conclusion*

This chapter set out with the lofty aim of tracking the influence the argument from increasing automation has on the design and autonomy of artificial moral agents. The force of this argument sprang from a number of separate concerns: the increasing automation of various task environments, the increasing autonomy of specific artificial agents, the entrance of AMAs into environments of ethical salience, and the worry that the convergence of these facts would lead to an unmitigated and unjustified risk of human harm. To understand this concern, we investigated the engineer's concept of autonomy, and the philosopher's concern for this autonomy in the first section. The engineer's concept of autonomy turned out to be a function of the agent's robustness to environmental change, where robustness was achieved without direct human intervention or supervision. Furthermore, there were two senses of the engineer's concept of autonomy: first, autonomy as *provision*, wherein the human programmer was able to preempt all possible changes in the agent's environment, and incorporate those possibilities into the AA's agent program, and second, autonomy as *independence*, where the agent itself learns new facts and features of its

environment, and is able to move beyond the *a priori* knowledge provided by the programmer. We also briefly maintained that these two visions of machine autonomy yielded two separate ethical concerns, if we could accept that the achievement of the agent's goals (the maximization of its performance measure) was a normative affair. From autonomy as provision, we understood that the designer is tasked with accurately describing what a machine most ought to do, given dynamic and changing environments. From autonomy as independence, we concluded that the machine itself may decide what it most ought to do, given dynamic and changing environments. In environments of moral salience, these visions both seemed to be cause for moral concern.

The philosopher's concern for machine autonomy, in turn, appeared to be conventionally tethered to the possible ascription of moral status to artificial agents, and the elucidation of what sorts of metaphysically thick capacities would need to be implemented for 'true' moral agency to be achieved. We offered the standard account of moral agency, the most popular approach in machine ethics, which relied on the existence of important mental states such as intentionality, desires and beliefs, and often presupposed consciousness, subjectivity and free will or moral autonomy. This view was challenged by the functional perspective of moral agency, which relied on levels of abstractions, empirical studies, or Turing tests to point towards the *appearance* of moral agency, which was often seen as sufficient grounds for its attribution. We observed three effects of the philosopher's concern for machine autonomy on the development of artificial moral agents: first, it exerted a *downward* pressure which advocated for the exclusion of artificial agents from environments of moral salience. Second, it exerted an *upward* pressure on the design of artificial moral agents, where design was encouraged to approximate the standard view of moral agency or other anthropomorphic features. Finally, it made the standard view and the humans who satisfy its conditions the paramount point of comparison across the various types of AMAs which could be conceivably built.

In that spirit, the second section investigated 6 types of potential artificial moral agent, ordered in increasing degrees of machine autonomy, as is often done in the literature. Levels 1 and 2 denoted ethical impact agents and implicit ethical agents, whose autonomy was minimal, and who were subject to a morality of use and a morality of design, respectively. Next, we investigated levels 3 and 4, which yielded two AMAs of particular interest: level 3 explicit ethical agents, and

level 4 explicit ethical agents. Ethical agents themselves were seen to be agents endowed with explicit normative programming which allowed them to respond to the moral value of their environment. We maintained that they were able to do this with relative independence, but that this was a feature of their artificial morality, and not their machine autonomy, as we had defined it. Both types of explicit artificial agent were subject to a morality of *behavior*, since their actions and decisions themselves could be subject to moral appraisal. What differentiated level 3 from level 4 was the capacity for *task autonomy*, or an independence over the means by which an agent achieves the goals posited by the programmer or user. Importantly, the morality of these agents was therefore tethered to the human programmer, they did not decide for themselves, in other words, which principles to act on or which values to pursue. Finally, we explored Moor's Brightline, and levels 5 and 6 of the scale of machine autonomy. Moor's bright line, as it turned out, touched on two closely related issues: the incarnation of the standard view in an artificial moral agent, and the critical threshold of moral responsibility. The problems of the attribution of moral responsibility in AMAs were seen to be plentiful in the literature, depending on the conservatism of our view. Conservatively, the so-called responsibility gap covers level 5 autonomy, since just like human agents, any AMA at this level would possess free will (or less ambitiously), *domain specific responsible agency*, which would allow it to 'stand its ground' in the face of protest or moral disapprobation.

Finally, we explored the first blushes of the *better angels of our nature* argument, which seemed to turn the assumed moral authority of human moral agents on its head. This argument held that moral agency could be improved via implementation into artificial moral agents, since the robot would be free of any human failings that traditionally prevent efficient moral action. We maintained that the clearest example of this idea, one that lead to a *deference* to the moral authority of machines, was found in the highly imaginative work on superintelligence. However, other instantiations of the argument were possible and perhaps prevalent in the literature. Having understood the various possible schemas of power sharing between humans and robots in environments of moral salience, we will in the next chapter hone in on the standard relation which exists between level 4 explicit ethical agents and human programmers or users, a schema which we will employ and limit ourselves to for the rest of the thesis.

CHAPTER

3

Heteronomy, Modularity &Artificial Moral Agents

In the previous two chapters, we have taken an in depth look at artificial agents, and how certain concerns for their autonomy in ethically salient environments have prompted the development of artificial *moral* agents. Throughout this process, we have witnessed a number of ways by which an artificial agent can be seen to differ from a human agent: an artificial agent is a technological artefact, not a natural entity; the design of an artificial agent is intentionally suited for action in a specific *Umwelt,* and not in social environments generally; and artificial moral agents, while performing actions for good or for evil, do not expound the conditions of the standard view of moral agency, they are not metaphysically thick persons with structural properties like consciousness, subjectivity or free will. Instead, as we have seen, most of what machine ethics considers to be artificial moral agents are *explicit ethical agents*, they mobilize explicit ethical content in their deliberations which allows them to respond to the moral value of their environment.

Still, in comparison to many of the more simplistic machines we have considered in the previous chapter, explicit ethical agents are privy to an impressive degree of independence, which,

in particular, manifests itself in their capacity for *task autonomy: "*...the ability of a system to adjust its behavior, by forming new plans to fulfil a goal, or by choosing between goals"[1]. We have characterized this idea of autonomy, in alignment with Dignum's original concept, as an autonomy of *means* over the *ends* provided by a human agent. In an autonomous vehicle, for example, the passenger indicates his desired destination, and the AV computes the most efficient route by which to reach it. This efficiency, in turn, may be affected by the autonomous vehicle's artificial morality: it may lead it to forgo certain routes, stop at certain distances, or if an unavoidable accident scenario is imminent, 'decide to crash' in certain ways. Clearly, the passenger, through employing an autonomous vehicle as a means to arrive at his desired destination, has relinquished quite a bit of his own autonomy. In effect, the passenger has delegated an important number of instrumental decisions over to the vehicle: he cannot choose—among other things—which route to pursue, how fast to drive, whether to stop for j-walkers, or perhaps, even to save his own life in an accident. In a highly autonomous AMA such as this driverless car, the only autonomy the human agent retains is his *goal autonomy*: what in engineering circles implies the ability to "…introduce new goals, modify existing goals and quit active goals"[2], but what is more commonly known in philosophical literature as simply *autonomy* or practical freedom. The human agent, in other words, is still afforded the freedom of acting on his own principles, higher values, or through the exercise of his own capacity for reason, to freely pursue the *ends* he pleases[3].

In the case of explicit ethical agents then, there exists something of a division of decisional labor: the human agent posits the *ends* the AMA is to pursue, and the AMA has an autonomy of *means* over the pursuit of these ends. This type of collective agency has precedence in philosophical literature, and is often referenced by the concept of *heteronomy*, which is perhaps most famously exposed by Emmanuel Kant. Roughly, Kant held that as human agents, in acting

[1] Dignum, 2019, 21.
[2] Dignum, 2019, 21.
[3] This particular description of autonomy aligns with Isaiah Berlin's conception of autonomy, which he characterizes as: "the wish on the part of the individual to be his own master. I wish my life and decisions to depend on myself, not on external forces of whatever kind. I wish to be the instrument of my own, not other men's acts of will. I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes which affect me, as it were, from the outside. I wish to be somebody, not nobody; a doer—deciding, not being decided for, self-directed and not acted upon by external nature or by other men as if I were a thing, or an animal, or a slave incapable of playing a human role, that is, of conceiving goals and policies of my own and realizing them." (Berlin, 2017, 126).

only according to those maxims that can be consistently willed as a universal law, we are not only determining our actions freely or autonomously, but are acting on our own self-determined law. Autonomous agents for Kant then, are those who act on their own law, or pure practical reason. Those whose practical agency is determined by something *external* to the faculty of practical reason however, are said to be *heteronomous*. From such a strict condition then, it follows that heteronomy can come in many forms. Kant, for instance, writes of the heteronomous employee who exercises his reason to decide the best way of achieving the ends laid out for him by his autonomous employer[4], other examples include the relationship between parents and children, king and subject, or even the relationship between chocolate and chocolate addict. In other words, the presence of any external force which influences one's faculty of practical reason—even internally—relegates one from autonomous to *heteronomous* agent.

Quite clearly, explicit ethical agents, such as those on the 4th level of machine autonomy, can be accurately described as heteronomous agents. Not only do these agents lack the goal autonomy indicative of the standard view of human agency (and thus, characteristic of human moral agents), but further, the types of moral principles and values that drive their decision-making in ethically salient environments are *human given*, a function of the artificial morality that the designer has chosen to implement[5]. In other words, two external forces exert influence over the artificial moral agent's practical deliberations: the goals or ends posited by the human agent, and the artificial morality imposed by the human programmer. To this end, Aristotle echoes this particular instantiation of the concept of heteronomy when he reflects upon the status of a slave as an "…instrument for instruments"[6]. Further, he characterizes the relationship between instruments, slaves and the master that presides over them as representing one, extended somatic system,

[4] Kant, 2013, 8:38

[5] We should clarify that the concept of heteronomy typically implies an underlying *capacity* for autonomy which is thwarted or constrained by the force of external influence. Since human beings are essentially autonomous, it is by force or by choice that they become heteronomous agents. The case of explicit ethical agents, or for that matter, artificial agents generally is decidedly different, since they cannot be seen to exhibit the type of metaphysically thick autonomy that Kant and other authors presuppose in their use of the concept. Thus, to be precise, explicit artificial agents are *necessarily* heteronomous, as a direct result of their type of design architecture and agent program. Hypothetically, machines which escape this necessity are those that sit on the 5th and 6th level of the scale of machine autonomy. Indeed, the goal of 'constraining' super intelligence to fit human moral standards can be seen as a deliberate attempt by humans to render an autonomous machine *heteronomous* to human morality.

[6] Aristotle, 1877, 1253b29-34.

"…since all instruments, animate or inanimate, human or not, are actually prosthetic extensions of the master, and as much a part of a master-centered network…"[7]. Thus, from an agentive point of view, explicit ethical agents become something of an exotic means for human ends, acting in intelligent an ethically sensitive ways in their *Umwelt*[8].

This concept of machine heteronomy garners further support from a seminal article written by James Moor, in which he questions whether there are some decisions 'that a machine shouldn't make'[9]. Initially, Moor draws a valuable distinction between the *ability* and the *authority* to make certain decisions: "…it might be argued that it is not computers which make decisions but rather humans who *use* computers to make decisions. But this point confuses the *power* to make decisions with the *ability* to make decisions"[10]. Moor takes the example of the president of the United States, who is vested with the *power* to make presidential decisions, even if many other citizens likely hold the *ability* to make such decisions. Here we see a weak political example of the classical concept of heteronomy: the president posits certain ends or elucidates certain principles in virtue of his authority, which come to impact or influence the practical decision-making of otherwise autonomous human citizens, rendering them heteronomous to his policies and projects. Initially, Moor defends that the situation between president and citizen is analogous to that of human and machine: "…we can delegate decision making to computers, and the fact that we use computers in this way is compatible with computers being decision-makers…To delegate decision-making power is to delegate control. Ultimately, the issue is what aspects of our lives, if any, computers should control"[11]. Thus the functional parity between human and artificial decision-makers is confirmed by our very employment of machines in our practical agency, and it is then something of a normative and authoritative choice to decide the scope of their decision-making.

Moor convincingly mobilizes this premise to attack his three, so-called *dubious maxims* of machine decision-making: that computers should never make any decisions which humans want

[7] LaGrandeur, 2013, 10.
[8] This idea is echoed in Martin Heidegger's vision of technology: "For to posit ends and procure and utilize the means to them is a human activity. The manufacture and utilization of equipment, tools and machines…and all the needs and ends they serve, all belong to what technology is" (1977, 3).
[9] Moor, 1979.
[10] Ibid, p. 219.
[11] Ibid, p. 219.

to make, that computers should never make any decisions which humans can make more competently, and that computers should never make any decisions which humans cannot override[12]. Since he assumes functional parity in decision-making between humans and machines, he is able to counter these maxims by appealing to the gains in efficiency which would likely result if these maxims were not categorically respected. Interestingly however, his arguments take on a moral tone. For instance, Moor claims that machines should make decisions humans want to make because "…there can be other factors which outweigh the benefits of the freedom and pleasure humans derive from doing decision-making"[13]; or that "…there could be situations in which it would be morally better to make it impossible, at least practically speaking, for humans to override computer decisions"[14], taking the retrospectively prophetic example of robotic cars and their promise of a reduction in traffic fatalities. We can thus see a connection with the technical form of the *better angels of our nature* argument of the previous chapter: if the practical agency of humans is seen to be imperfect or is seen to cause unnecessary harm, and artificial agents are immune to these failings, then we *ought* to delegate decisional control to artificial agents.

Moor's tone changes, however, in the conclusion of his argument, where he questions whether there are any areas of human practical agency which machines should never control. His answer is telling:

> Computers should never decide what our basic goals and values (and priorities among them) should be. These basic goals and values, such as the promotion of human life and happiness, decrease in suffering, search for truth and understanding etc., provide us with the ultimate norms for directing and judging actions and decision-making. By definition there are not further goals and values by which to evaluate these. Since we want computers to work for our ends, we obviously want to prohibit computers from deciding to change these ultimate norms…

Clearly, Moor's thoughts fall within the bounds of the means-ends type of reasoning which is indicative of the concept of heteronomy. Interestingly, while he advocates a functional view of machine agency—one which ignores the problem of a machine's lack of consciousness, free will, or its failure to espouse the standard view—he nevertheless maintains that machines ought to be

[12] Ibid., p 226-227.
[13] Ibid., p 226.
[14] Moor, 1979, 227.

heteronomous to human ends and values. They ought, in other words, never to hold the *authority* to decide humanity's basic goals and values, even if they have such a capacity.

Thus, the concept of heteronomy sits at an interesting intersection in machine ethics: the standard view holds that machines cannot be autonomous moral agents, since they lack the requisite metaphysically thick structural properties. They are, from this perspective then, *necessarily heteronomous* to human ends, which is typified by the *autonomy of ends* humans hold over machine decision-making, and the *artificial morality* human programmers implement into the machine, allowing them to be explicit ethical agents who are responsive to moral value. From the functionalist perspective, however, there is a *parity* between the decision-making capacity of humans and machines. Nevertheless, there are strong, normative reasons to retain the *authority* over the ends which the machine can pursue. For the functionalist then, machines are not necessarily heteronomous to human ends and value, but they nevertheless, from a humanistic perspective, *ought not to make* these sorts of decisions[15]. Thus, while the standard and functional view may offer different arguments towards machine heteronomy—a lack of capacity or a lack of authority, respectively—both support the view that goal autonomy, and the choice of the principles, values and reasons which motivate machine action, ought to be reserved to human agents.

In a broader sense then, we can see how these two perspectives, and the argument from increasing automation, converge on the design of level 4 explicit ethical agents. Indeed, it is only these agents who are a) involved in morally salient environments where the risk of human harm is prevalent, b) endowed with a capacity for task autonomy, or an *autonomy of means* in their practical agency, c) are equipped with a human-given artificial morality which constrains their move from perception to morally responsive action within their environment, and correspondingly, d) are deprived of goal autonomy. Anecdotally, we may also perceive a certain temptation, from the better angels of our nature argument, to delegate significant portions of the *means* by which we achieve human ends over to machines—for instance to maximize human safety or positive

[15] Moor's justification for this particular claim is precautionary in nature: "…we obviously want to prohibit computers from deciding to change these ultimate norms, e.g., promoting computer welfare at the expense of human welfare or taking inconsistency to be the mark of good reasoning" (1979, 227).

moral impact— adopting something of a maximalist approach to the design of heteronomous machines[16].

Together then, these pressures provide the essential frame or limits which surround the practical agency of artificial moral agents, and our analysis will accordingly home in on level 4 explicit ethical agents which expound these criteria. Henceforth, we will consider only these machines to be artificial moral agents or refer to them as such. The goal of this chapter will be to specify, in precise terms, how the condition of these agents differs from that of a typical human agent, and to understand two basic structures of power-sharing between humans and machines. In section I, we will explore the concept of *modularity,* which we will use to describe the truncated type of moral agency to which AMAs are necessarily privy. In section II, we will investigate two different power-sharing structures, yielding two different types of agent: what we will call surrogate agents, and what we will call distributive agents.

1. *Modular Artificial Moral Agents*

In 1951, Paul Fitts introduced what has come to be known alternatively as the HABA-MABA framework or the 'Fitts List' for function allocation in sociotechnical systems[17]. In the report, Fitts listed 11 functions which 'Humans Are—or 'Machines Are'—Better At'; where human beings were seen to excel in such areas as perception, judgement, induction or improvisation; and machines: speed, power, computation and replication, among others. Despite the era in which it was written, Fitts' list appears to have aged gracefully, and remains a widely consulted paper in allocations research[18]. Of course, while it remains certainly true, in the abstract, that machines are better at speed and computation than human beings, this cannot be true of every instantiation of a machine. A machine which is *designed* to compute—which is to say, its purpose-oriented ontology includes the design goal of computation—may certainly compute better than a human agent, but a machine which is designed to sort different meats into the appropriate bins might not. In a similar vein, a given machine may be stronger, or move faster than a human agent,

[16] This particular temptation, which plays out in the design of an AMA's artificial morality, will occupy us at great length in part II of this thesis.
[17] Fitts, 1951.
[18] de Winter & Dodou, 2014.

but *every* machine cannot be said to excel in this way. The inverse, however, is not true of (most) human agents. A single human being, say a trombone player in a high school orchestra, could likely outstrip machines not only in (jazz) improvisation, but also perception, judgement and induction, while certainly not being an expert in any of these respects[19]. This asymmetry points to an essential ontological difference between human and artificial agents, namely, that the former are *universal* practical agents, while the latter are *modular*: evolving and interacting within a specific environment, or *Umwelt.*

Specifically, the modularity of artificial moral agents implies three different types of limits on their practical agency: the *goals* the agent is able to pursue (extensional restrictions), the types of agents the AMA is able to perceive (agentive restrictions), and the types of decisions it is able to make (intentional restrictions). We will illustrate these distinctions through the analysis of a highly fictitious artificial moral agent, BerryPicker3000™.

> **BerryPicker3000™**: One fine day, Bob, a designer at Baltimore Dynamics, comes up with a new idea for an artificial agent, BerryPicker3000™. This embodied artificial agent is designed to pick berries from specific types of vegetation: blackberry, blueberry and raspberry bushes. Accordingly, Bob designs BerryPicker3000™ to pick *only* the berries from these bushes, by providing world knowledge about the appearance of their foliage, and by distinguishing these appropriate bushes from other varieties that BerryPicker3000™ is forbidden to pick: cherries, boysenberries, and strawberries. Finally, Bob designs an order of priority for optimal berry-picking: blueberries, raspberries and blackberries, since the softer berries are liable to be crushed if picked first. Needless to say, the suits at Baltimore Dynamics are thrilled, and green light BerryPicker3000™ for immediate production.

As a point of departure, we can easily make a connection between the purpose of this artificial agent and its environment: BerryPicker3000™ is designed to pick berries, and its context of action (or its *Umwelt*) therefore plausibly includes contexts like agricultural fields and personal gardens. Thus, to frame BerryPicker3000™ within our definition of artificial agents from the first chapter, we can say that BerryPicker3000™ is a technological artefact capable of flexible action

[19] While modern AI is still unable to compose a convincing jazz song, DARPA has recently invested their time and attention into teaching it how to 'jam', with relatively dismal results : https://youtu.be/O-bjTfYILPs

within agricultural and personal berry gardens, in order to meet its design objective: to pick berries. The *extensional* restriction then, involves the purpose or design objective of the artificial agent. Indeed, while BerryPicker3000™ is endowed with an autonomy of means over *how* it picks berries, it cannot decide to quit berry picking for good; in pursuit of, say, a career in flower picking or meat rendering[20]. Further, what we have called the *purpose-oriented ontology* of BerryPicker3000™ is connected to this extensional restriction. In other words, in order for this agent to effectively pick berries, it might require a specific type of agent program, and specific actuators and sensors. Finally, the extensional restrictions are not only directly reducible to the design objective of a given artificial agent, they also indicate over which area the machine is seen to possess an autonomy of means. Intuitively, this is true because any human agent who elects to employ an artificial agent as a means to achieve his ends, likely does so in function of the posited purpose of the machine. A rational human agent, in other words, would not use or purchase BerryPicker3000™ for anything other than berry collection, or other more long-term goals such as pie-baking, jam-making, or increased agricultural efficiency and profit.

Second, we can observe that BerryPicker3000™ is designed to pick berries from some types of vegetation (blackberry, blueberry and raspberry bushes) but not others (cherry, boysenberry and strawberry bushes). These relate to the *agentive restrictions* of the artificial agent, and serve to limit the types of phenomenal signature an artificial agent can perceive in its *Umwelt*. In terms of our definition of artificial agents, the agentive restrictions impact the agent's capacity for both reactivity and proactivity. This is because the phenomenal signatures that are encompassed by the agentive restrictions are tantamount to the *marks of significance* within the agent's environment. In other words, for BerryPicker3000™ only *some* berries are significant or seen to have value: blackberries, blueberries and raspberries. The agent's world knowledge then, would pertain to these types of berries. For instance, Bob the designer, in intentionally designing BerryPicker3000™ to be responsive to these types of berries, likely provided specific *a priori* knowledge about these entities, such as the fact that blueberries are blue, blackberries have thorns, and perhaps that low-hanging berries of any kind are susceptible to animal contamination. In this way, BerryPicker3000™'s capacity for reactivity reduces to its ability to perceive blackberries,

[20] This would be a rather lyrical example of goal autonomy, implying the autonomous action of this AMA.

blueberries and raspberries, to respond to changes that occur within them (such as when they are under or over-ripe), and possibly to learn how best to adopt these changes, mobilizing relevant facts such as the idea that blackberries have thorns. Its capacity for proactivity in turn, relates to its ability to pursue internal goals in a way that is responsive to these types of berries: it may choose to pick raspberries first if they are seen to be more ripe, or refrain from picking blackberries on certain days if over-picking would result in a decrease in berry production. Depending on the type of architecture Bob the designer implemented, BerryPicker3000™ may learn new things about berry picking which would render it more efficient (autonomy as independence), or Bob may have been very thorough in his prediction of all possible changes within the berry picking environment (autonomy as provision). Thus, the agentive restrictions serve to define which of the various entities in the agent's environment are seen to count as marks of significance, and accordingly, the agent must be responsive to the value of these entities. Of course, in the case of BerryPicker3000™, it may seem strange to consider berries as 'agents' or as having 'value', but what is true of berries in our current example is true of pedestrians and other vehicles in the case of autonomous vehicles, or military targets and non-combatants in the case of autonomous weapons. In more general terms then, the agentive restrictions define the entities which are seen to have a *moral value* to which the AMA must respond.

Finally, we can see that Bob the designer has devised a particular strategy for BerryPicker3000™'s berry picking: to pick the softest, or most structurally vulnerable berries last, so as to avoid crushing them from the weight of the rest of BerryPicker3000™'s daily harvest. To simplify, we could reformulate this strategy in the form of a rule or maxim: 'pick the sturdiest berries first'. Of course, this is but one of many possible berry-picking strategies Bob the designer could have chosen. For instance, he could have also implemented the berry-picking maxim 'blueberries first', or 'blackberries last' to similar effect. In any case, the strategy chosen to inform BerryPicker3000™'s practical behavior, telling it—so to speak—how to accomplish its goal, encompasses its *intentional restrictions.* There are many factors which can inform the intentional restrictions of an artificial agent: the maximization of efficiency, the minimization of waste or damage, the catering to a user's preference, or adherence to overarching norms, legal standards or general acceptability concerns. In this way, the intentional restrictions are an integral part of an AMA's artificial morality. This is because the intentional restrictions have a considerable impact

on the means by which an AMA achieves its goal, precluding it from pursuing some (ethically sub-optimal or forbidden) options in favor of other (permissible or optimal) options. In this sense, the intentional restrictions of an artificial agent affect its capacity for proactivity and sociability. These restrictions preclude or forbid certain internal goals and possible actions in the agent's environment, and they may in turn affect the means by which the AA interacts with its environment.

At this point, it should be appropriate to clarify the interaction between the modularity of artificial agents as we have just described it, the concept of machine heteronomy, and the concept of artificial morality. From the concept of heteronomy, we can glean two key points: first, that the AMA is seen to have an autonomy of means over a given human agent's ends, and second, that the implementation of explicit ethical maxims, constraints or decision systems—those that flow from the status of AMAs as explicit ethical agents—is accomplished by the human programmer. Put simply then, the concept of machine heteronomy suggests that any AMA is 'tethered' both to the *ends* and to the *moral principles or policies* of one or more human agents. From the concept of modularity, we first have the general idea that the practical agency of an artificial agent is truncated (or modular) compared to the universal agency of a typical human agent; the machine is designed *for*, and an efficient agent *in*, a specific *Umwelt*, and not in social contexts generally. From this point, there are three specific limitations the designer can place on its (moral) agency. First, extensional restrictions, which underscore modularity by limiting the goals the machine can pursue to those that align with its purpose in a given *Umwelt*. Second, agentive restrictions, which render the agent sensitive to only a *specific set of phenomenal signatures* in its environment, those which are deemed to have (moral) value. This is not to say that the agent is 'blind' to entities other than those which can count as marks of significance in its environment, but rather that they are 'neutral' or adiaphoric, and therefore do not require responsiveness from the agent. Finally, the intentional restrictions serve to limit the decisional aspects of the machine: what sorts of values, principles or rules its actions can be seen to respect or uphold, and what sorts of actions can count as optimal, permissible or forbidden. Artificial morality then, encompasses both the agentive and intentional restrictions of the AMA; since together, these restrictions describe *what is of value* in a machine's environment, and *how it ought to respond to this value*.

In a clear sense then, the concepts of modularity, heteronomy and artificial morality align with what has often been called 'bounded morality'[21] or 'limited-domain robotics'[22] in the literature. Indeed, Wallach & Allen see bounded morality as a necessity in explicit ethical agents, since "…Just as it would be dangerous to put a chain saw in the hands of a child or the hands of an adult who had no training in its use, so too would placing a robot in a context where it would encounter challenges it neither recognized nor had the means for determining what actions were safe and appropriate"[23]. Here again, we can detect the precautionary premise of the argument from increasing automation: in the absence of full ethical agents (either human or machine), bounded morality is necessary to ensure that human harm will not result from a robot's actions and decisions. In ways complimentary to our analysis, Wallach & Allen define bounded morality in terms of the agent's intelligence, its sensitivity to its context, and what they call the 'ethical routines' it has to determine which actions are morally acceptable—which we might tightly understand to encompass the intentional restrictions, or more broadly, artificial morality[24].

2. *Surrogate Agents and Distributive Agents*

Just as the practical (moral) agency of an AMA is seen to be bounded or modular compared to that of a universal human agent, we can draw a loosely analogous distinction within the *mobilization* of moral principles, maxims rules and values between the deliberation of humans and machines. The details of this distinction will occupy us at great length in the second section of this thesis, but we will introduce our foundational points here.

[21] "Bounded morality refers to adhering to moral standards within the situations that a system has been designed for…and not in a more general sense" (Arkin, 2015, 46).
[22] Abney, 2012.
[23] Wallach & Allen, 2012, 127.
[24] "When designers and engineers cannot fully anticipate when and where a functionally moral robot will encounter a challenge they will need to understand: (1) the space (the environment) in which the robot operates well enough to ensure that the system recognizes when it is in an ethically significant situation. (2) The routines the system will require for determining an appropriate course of action….The bounded morality of a robot will be structured by its intelligence, that is, by its sensitivity to features and changes within that context, and by the ethical routines it has for determining which actions are morally acceptable within that situation." (Wallach & Allen, 2012, 127).

The principles which drive a *human* agent's responsiveness to moral value—whether they be autonomously or heteronomously derived—tend to apply universally, at least for the non-casuistic traditions of moral philosophy. If a given human agent is motivated by utilitarian principles, for instance, then these principles ought likely to matter in any context of moral salience in which the agent could find himself. This means that regardless of whether this agent is deciding which career to pursue, which charity to support, or whether or not to use expired eggs when cooking his wife's omelette, his moral principles will lead him to choose that action which brings about the greatest happiness. In this way, there is a general pressure in traditional moral theory towards the establishment of universal principles; action-guiding recommendations or permissibility verdicts which can be seen to apply anywhere across a human agent's universal practical agency.

The case of machines, on the other hand, is somewhat different. While a designer may choose to implement certain universal principles into the artificial morality of an AMA, this choice may have less to do with the scope of their applicability, and more to do with the type of response to moral value they could be seen to yield within a given *Umwelt.* We can see this basic trend in machine ethics: of the various ethical doctrines in theoretical and applied philosophy certain schools are seen to apply more readily to certain types of AMAs. Specialists in the design of robotic healthcare assistants, for instance, often espouse an artificial morality which takes inspiration from the ethics of care tradition[25], while specialists in the field of autonomous weapons—notably Ronald Arkin—often support an artificial morality which aligns with Just War Theory or the Rules of Engagement[26]. Still, this tendency to design an agent's artificial morality around more 'local' normative standards rather than universal maxims is not borne out by all of the world's machine ethicists. As we shall see, in both the early years of machine ethics and quite recently, a large number of specialists have attempted to design 'Kantian', 'Smithian' or 'Rawlsian' machines which take a more universal approach to the design of artificial morality. In broad terms however, there seems to be a trend towards the specification of moral responsiveness to the features of an artificial agent's prospective *Umwelt.*

[25] van Wynsberghe, 2012: 2013: 2016: Stahl & Coeckelbergh, 2016.
[26] Arkin, 2009: Sparrow, 2007: Wallach & Allen, 2012.

To this end, one often overlooked corner in general machine ethics literature relates to the relationship between the role of an AMA vis a vis *a particular user* and the type of artificial morality which may be appropriate to implement. For instance, one somewhat insular cluster in the machine ethics literature relates to the ethics of human-robot interaction (HRI) or 'social robotics', a field which puts particular emphasis not only on the safety of such agents, but also on the development of bonds of trust between such agents and their principal user[27]. This may pass through the simulation of emotions and the robot's capacity for emotion detection, but also through the implication and control the principal user has over the agent's behavior[28]. On the other side of the spectrum, considerable attention has been given to the ethics of machines who are seen to interact with *multiple* users, none of whom are of singular importance to the machine. These include expert systems of various kinds, such as autonomous vehicles and weapons, but also some types of robot healthcare workers and many virtual artificial agents. This division across the discipline of machine ethics shares interesting common ground with a long-assumed but often neglected trend in the field: the conception of artificial agents as *surrogates* or *proxies* for specific human agents.

From a moral point of view, there may be some salt to the idea that certain robots, who occupy certain roles in society, may be morally required, or at least expected to behave in special ways to their users. In this spirit, we will posit a distinction between two *types* of artificial moral agent: *surrogate* agents, who are seen to act in the interest of or on *behalf* of a specific human agent (typically referred to as the 'user'), and what we will call *distributive* agents, who bear no particular attachments to the users with whom they interact. While these are hardly clear-cut categories, we would do well to explore the differences between them, so as to begin to understand their influences over the appropriate design of artificial morality.

2.1 *Surrogate Agents*

In the early to middle years of the field of machine ethics, many authors viewed artificial agents as technological artefacts who performed actions on behalf of, or in the interest of, particular

[27] Salem et al., 2015 : Brezeal, Dautenhahn & Kanda, 2016.
[28] Salem et al., 2015.

human agents. Indeed, this was even seen to be a foundational aspect of the concept of an artificial agent[29]. Intuitively, the popularity of this view was likely relative to the types of artificial agents which were the most common object of analysis at the time: software agents or minimally autonomous embodied agents who performed specific tasks for users. To this end, perhaps the most thorough espousal of the surrogate view of artificial agents can be found in Johnson & Powers' article, *Computers as Surrogate Agents*[30]. Their argument revolves around a central, strong claim: "…computer systems, like human surrogate agents, perform tasks on behalf of users. They implement actions in pursuit of the interests of others. As a user interacts with a computer system, the system achieves some of the user's ends"[31]. Through an argument by analogy, Johnson & Powers are able to develop a robust moral framework which centers around a particular concept of surrogacy: an agent who adopts a third-person perspective and pursues the 2nd order interest or desires of his client. Tax attorneys, lawyers or medical experts, they claim, all operate within this framework, and all are subject to moral evaluation in so far as "…the agent is incompetent or misbehaves with respect to the client's interest"[32]. Incompetence they claim, relates to the inability of the agent to adequately perform its task, due to internal failings or the complexity of the world. This is cause for moral concern precisely because these agents are unable to perform the job they were ostensibly 'hired' to do, such as what happens when a tax attorney fails to submit his client's taxes on time. Misbehavior, in turn, relates to the surrogate's pursuit of the interest of someone *other* than the client, typically in ways which are detrimental to the client[33]. A defense attorney who makes arguments for the prosecution is an unfortunate example of misbehavior.

[29] To this end, we might recall the third definition of artificial agents from the first chapter of this thesis: "Autonomous agents are software entities that carry out some set of operations on behalf of a user or another program with some degree of independence or autonomy, and in so doing, employ some knowledge or representation of the user's goals of desires"(Franklin & Graesser, 1996, 23). In a similar vein, John Sullins defines artificial agents as "…any technology created to act as an agent, either as a locus of its own power, or as a proxy acting on behalf of another agent. So an artificial agent might have its own goals that it attempts to advance, or more likely, it is created to advance the goals of some other agent" (2009, 206). Other espousals of the idea of surrogate agents can be found in Millar, 2014a: 2015: Sandberg & Bradshaw-Martin, 2013: Dignum, 2019.
[30] Johnson & Powers, 2008.
[31] Ibid., p. 257.
[32] Ibid., p. 260.
[33] Ibid., p. 260.

Johnson & Powers thus provide convincing grounds for the claim that the purpose-oriented ontology of an artificial agent, as well as its extensional restrictions, may paint their behavior as *morally* tied to a particular human user and his interest, and that they may be subject to moral disapprobation when they fail to expound this interest through incompetence, or through the pursuit of other aims. This conception opens us up to one key idea: that it may be morally permissible or perhaps even required that artificial agents act on *restricted reasons* in their ethical deliberation. This would entail the claim that an AA would be justified in "…adopting the partial aims and point of view of the partisan, thereby restricting the range of moral reasons that count in one's deliberations, so that some good moral reasons are excluded or discounted, and others are given priority or magnified…"[34]. In more simple terms, this claim points to the potential for morally admirable partiality in the moral behavior of artificial moral agents, a capacity to include agent-relative reasons for action in the machine's artificial morality. Additionally, Johnson & Powers' view of surrogacy may require an added constraint on the behavior of an AMA: that it should never intentionally thwart the user's interest in pursuit of another aim or end, thereby avoiding the charge of misbehavior.

Under this view then, the concept of machine heteronomy is taken rather narrowly. In effect, the perception of artificial agents as surrogate agents would entail that the agent has an autonomy of means over *exclusively* user-centric ends. Furthermore, the type of artificial morality which would be morally appropriate to implement into surrogate agents would likewise be highly user-centric: either an espousal of the moral principles and preferences of the user, or in strict alignment with Johnson & Powers, a rational framework which would guarantee the pursuit of the user's second-order interest. Of course, this is a somewhat severe form of the concept of surrogate agents, since it leaves open the possibility of a human user employing an artificial agent to intentionally harm a human being, such as what might occur if an autonomous vehicle were able to target or kill individual pedestrians at the user's behest. This seems neither desirable nor morally sound, if only because it would fail to respect the precautionary concern of the argument from

[34] Applbaum, 1999, 5.

increasing automation[35]. If the designers of artificial agents left open such a possibility, they could not ensure that human harm would not occur as a result of the machine's behavior.

To this end, we can adopt a softer, more general conception of surrogate agents. Generally then, the design of a surrogate agent must meet the following conditions:

a) the design and purpose of an artificial agent must entail frequent and prolonged interaction with a specific user
b) The agent must have an autonomy of means over this user's ends, acting on behalf or in the interest of this user within its *Umwelt*

There are many artificial agents which ostensibly meet these requirements. Digital assistants such as Alexa or Siri, robot vacuums, and a slew of more imaginative machines who may one day perform tasks on behalf of their users: a robot that fetches groceries for a person, or a robot who dispenses medication and monitors an elderly person. Essential to the concept of surrogate agency is the idea of *morally admirable partiality*: the machine may minimally be permitted to act on restricted reasons, and in so doing, privilege the interest or welfare of its user over that of other human agents. And maximally, it may be required to privilege the interest or welfare of its user at the *expense* of other human agents, while stopping short of engaging in an act of intentional harm. For the design of artificial morality then, surrogate agents provide the grounds to include partial, agent-relative responses to the moral value of the agent's environment.

2.2 *Distributive Agents*

Despite the soft tradition of the surrogate conception of AMAs in machine ethics, it is clear that the brunt of artificial agents which are currently the objects of analysis and inquiry are not seen to hold any special relationship with a given user. Indeed, often times, if they are tools or assistants at all, they are assisting in collective, societal aims, such as the reduction of traffic fatalities through the implementation of autonomous vehicles, or the reduction of casualties in

[35] For that matter, it fails to respect perhaps the ultimate tenet in moral philosophy, one which prohibits the act of intentionally harming a human being, often called the *harm principle* (Feinberg, 1987).

autonomous weapons. For this reason, the types of artificial morality proposed for these types of literature has taken on something of an *impersonal* character: utilitarian[36] and Rawlsian[37] approaches to autonomous vehicle decision-making being two such examples. In other words, these agents are conceived as something akin to independent actors, who *distribute* their services (or when necessary, the harm they can inflict) across various human agents of equal importance. An autonomous weapon, for instance, should not spare the lives of some enemy combatants over others; and an autonomous vehicle should not privilege the lives of certain members of society (say, priests or bankers) over others. Thus, the types of artificial morality typically proposed for distributive agents often have strong ties to the ideals of equality and fairness. Indeed, when a distributive agent's decision-making procedure fails to be equal and fair, objections such as machine bias are often levelled against it, as was the case with the COMPAS system's regrettable tendency to predict a higher recidivism rate for black offenders[38].

We can restate this tendency more precisely by claiming that distributive agents are often expected to expound *impartial* or agent-neutral responses to moral value across their actions in a given *Umwelt*. Importantly, this drive towards objective, impartial reasons shares strong attachments with the idea of 'AI for good', or the view that artificial agents ought to promote collective ends and benefits across their implementation and subsequent decision-making. Indeed, in the absence of a clear principal user, we would be hard pressed to find a (morally acceptable) reason for an artificial agent to be partial to a given human agent's preferences or claims[39].

Thus, we can identify the conditions of distributive agents as follows:

[36] Bonnefon et al., 2016.
[37] Leben, 2017.
[38] Perry, 2013 : Dressel & Farid, 2018.
[39] This idea generates a part of its force from the concern many have over the types of private interests that certain, especially virtual artificial agents could pursue when acting in society, a long-standing concern in the field of Technoethics (Bunge, 1977). In this sense, even if an artificial agent is ostensibly acting on behalf of a given corporate entity, and in this sense may appear to qualify as a surrogate agent, the interaction and impact these technologies will have on society preclude it from acting on partial reasons.

a) the design and purpose of an artificial agent must entail frequent and prolonged interaction with multiple users, none of which hold any special relations to the machine
b) The agent must have an autonomy of means over general, or collective ends, acting on behalf of no particular entity in its *Umwelt*.

Of course, it is superficially true that virtually every artificial agent, in virtue of its being designed by a particular human agent within the larger structure of a sociotechnical system, is often *representing* a certain entity across its behavior. For instance, Tesla's persistent problems with its Autopilot technology may damage this company's *reputation*, but we would be hard pressed indeed to claim that this recalcitrant failure to avoid rear-ending vehicles is serving the *interest* of Tesla or its engineers. Thus, it may be the case that these vehicles are serving Tesla's aims very broadly construed (through the domination of the driverless car commercial market for example), but not in a way that is sufficiently precise to trigger the concept of surrogacy. Hopefully, Tesla's Autopilot technology aims at the collective benefit of a reduction in traffic fatalities through the elimination of human error in the traffic environment, even if it currently fails to achieve this end consistently.

Importantly, it is often not obliquely apparent whether a given artificial agent pertains to the category of surrogate or distributive agent. After all, autonomous vehicles may 'distribute' the harm they cause in unavoidable accidents, but they may very likely bear certain obligations to their passengers, especially if these passengers have no control over the behavior of the machine, or if they own the vehicle[40]. This distinction is further complicated by the collective use of many artificial agents: each individual user may treat Alexa as a surrogate for their shopping needs, but every individual, collectively, interacts with Alexa (and the marketplace she represents) precisely for this purpose. Correspondingly, the types of principles, maxims, rules and values we might seek to implement into an AMA's artificial morality may shift depending on the way in which we view a given artificial agent, causing us to reject user-centric preferences and claims towards more impartial perspectives, or vice versa. In this way, the role that an artificial agent plays in society,

[40] Lin, 2016 : Millar et al., 2017 : Keeling et al., 2019 : Evans et al., 2020.

as a surrogate or as a distributive agent, plays a fundamental part in the design of its artificial morality, and has serious and lasting impacts on auxiliary concerns such as liability, acceptability and user trust. Tempting as it is to follow these leads further, we must reserve this discussion until our understanding of artificial morality has matured.

3. *Conclusion*

This chapter began with the elucidation of the concept of heteronomy: the external influence of an agent's practical agency, causing it to be motivated by certain principles, or to pursue certain ends which are not self-determined. We established that level 4 explicit ethical agents—those upon which our analysis will now focus—are heteronomous in two respects: in regard to the *ends* the agent pursues, and in regards to the policies, maxims, rules and values its actions reflect, both of which are posited by a human agent or programmer. Put simply, AMAs are heteronomous because of their purpose-oriented ontology, and their artificial morality. We likewise understood that there existed something of an informal consensus surrounding machine heteronomy in the literature: from the standard view, these agents do not have the *capacity* to be motivated by self-determined principles, and from the functionalist perspective, these agents do not have the *authority* to determine the ends for which their actions are a means. In both cases then, human agents were seen to be charged with the role of the 'designator of ends' for artificial moral agents.

We then investigated the concept of modularity, which was seen to capture the ways in which machines differ from universal practical (human) agents. Specifically, there were three ways in which machine agency was seen to be truncated: the extensional restrictions, which served to limit the types of general goals or purposes an artificial agent was able to pursue in its *Umwelt*, the agentive restrictions, which limited which types of entities could hold value or count as marks of significance in the agent's *Umwelt*, and thus to which entities it must respond morally, and finally intentional restrictions, which determined how the agent would respond to the value in its *Umwelt*. Thus, the concept of extensional restrictions aligns quite strictly with an agent's purpose-oriented ontology, while the agentive and intentional restrictions together define its artificial morality, or the *what* and the *how* of its moral responsiveness.

Finally, we explored two types of artificial moral agent: surrogate and distributive agents. The clearest distinction between the two relates to their relationship to a particular human agent: surrogate agents are seen to act in the interest of or on behalf of a particular human agent, while distributive agents hold no special relationships with particular human agents, and are in this sense, more 'independent' actors. Put another way, surrogate and distributive agents track two different applications of the concept of machine heteronomy. Surrogate agents clearly pursue the ends of the agent's principal user, and may thus be endowed with an artificial morality which admits the possibility of morally admirable partiality, or agent-relative, *partial* reasons for action. Distributive agents, on the other hand, can be loosely viewed as pursuing general ends, such as various collective aims or societal flourishing. In this sense, their artificial morality often admits the possibility of *impartial* or agent-neutral reasons for action, and it may be difficult to justify responses to moral value which are seen to favor or be biased towards specific members of a society. Importantly, the distinction between surrogate and distributive agents is not clear-cut: many real-life artificial agents may meet the conditions of both surrogate and distributive agents, depending on the perspective from which they are analyzed; moreover, many artificial agents may be subject to communal use, thus further blurring these boundaries.

Nevertheless, we emerge from this chapter with a clear view of the object of artificial morality: level 4 explicit ethical agents, or what we have described as heteronomous modular artificial moral agents (HMAMA). To recapitulate, these are the only non-biological agents who are: a) involved in morally salient *Umwelts* where the risk of human harm is possible or prevalent, b) endowed with a capacity for task autonomy, or an *autonomy of means* within their specific *Umwelt*, c) are equipped with an artificial morality which describes *what* is of value in the agent's Umwelt, and *how* it is of value, and d) are deprived of goal autonomy, or truly self-determined moral action. If agents of this kind are the 'protagonists' of the saga of machine ethics, then the argument from increasing automation, and more subtly, the better angels of our nature argument, together provide the inciting incident for its story. This is the case since, on the one hand, the moral relevance of a HMAMA may be evaluated in a precautionary, almost negative sense: to avoid the risk of human harm as a consequence of the agent's action, or to minimize harm and the production of 'unwanted events' which result from their implementation. But on the other hand, they also

seem to serve a positive role: through the implementation of HMAMAs, we may be able to *maximize* societal gains, correct for human failings, and make meaningful strides towards moral progress. The design challenge for HMAMAs, in this sense, is to navigate between and across these two terrains, specifically in the design of an agent's artificial morality. The next part of this thesis investigates how we might attempt this journey.

PART

II

Artificial Morality & The Hard Problem of Machine Ethics

In the last section, our principal aim was to uncover some of the basic ontological characteristics of what we have come to call *artificial moral agents*, those artificial agents who a) operate in contexts of ethical salience, and b) are equipped with a decision-making procedure which allows them to mobilize ethical principles, maxims or rules in their practical deliberations. We have called such decision procedures *artificial morality*. We have also maintained that the principal justification for the creation of this type of artificial agent hails from the *argument from increasing automation*: if moral value is an inherent feature of an artificial agent's environment, then a failure to account for this moral value in the agent's decision-making will result in the risk of human harm. Finally, we have identified artificial moral agents as entities which are both heteronomous and modular in comparison to human agents. They are heteronomous, since their practical agency serves as a means to the attainment of human ends, be they individual as in the case of *surrogate agents*, or common and general, as in the case of *distributive agents.* Their autonomy, whether it be human given (provision), or self-divined (independence) is then restricted to deliberation about these means. They are also modular, since their purpose-oriented ontology

admits three types of restrictions: *extensional* restrictions, which limit the practical agency of the AMA to tasks pertaining to its specific function or role, *intentional* restrictions, which limit the strategies it is able to employ in pursuing its practical goals, and *agentive* restrictions, which identify which entities are considered to be bearers of moral value in the AMA's environment. In all of these ways, artificial moral agents are the truncated, metaphysically thin counterparts of universal human moral agents.

In light of this analysis, it should seem plausible that the type of morality which is appropriate for these types of agents may differ from 'human morality' in important ways. The goal of part II will be to explore the extent of this difference, and its implications for the design of artificial moral agents. However, by way of preliminary remarks, it should be useful to revisit the concept of 'machine ethics', and how research from this field might influence our exploration of artificial morality.

Indeed in the literature and in the world, there is a problematic lack of consensus surrounding the meaning and purpose of machine ethics. For some authors, machine ethics resembles a branch of applied ethics. If this were true, the methodology of machine ethics would then have much in common with what is done in business ethics or medical ethics: the application of moral concepts and paradigms to a given institutional practice, with the end of providing an evaluative or prescriptive paradigm. Understood this way, machine ethics is circumscribed by the institutional practice of engineering, or more specifically perhaps, the research and development of artificial intelligence. In this sense, a machine ethicist is someone who makes evaluative or prescriptive claims about the deontology of designers, engineers and AI practitioners. The leading question of this vision of machine ethics is: how to ensure that engineers act on the right principles (or embody the right type of moral agent) in the design of artificial agents?[1]

A second, equally popular way to define machine ethics revolves around a rather theoretical question: how to design an artificial agent that reasons (or acts) like a moral agent? For these authors, machine ethics resembles an exercise in moral philosophy, and therefore abides by

[1] This is equatable to what we have called in previous chapters 'ethics for designers' (Dignum, 2019, 6-7), or the first sense of Verrugio's (2005) appraisal of machine ethics.

the rules and practices of theoretical engagement and debate. In this sense, a machine ethicist is someone who makes an argumentative contribution to a larger conversation concerning such topics as: whether artificial agents could be moral agents or patients, whether there are any moral 'bright lines' in the design of AMAs, whether a given moral theory is appropriate for implementation in robotics, how an artificial moral agent ought to reason, and *what* ought to matter morally to a machine[2]. Often, the quality of these contributions is judged on *rational* grounds—i.e. internal coherence and argumentative clarity—rather than in more practical or pragmatic terms—i.e. technical feasibility or desirability. This leaves open the possibility that machine ethics can produce arguments of astonishing insight and coherence, while remaining entirely unfit for implementation in current, and perhaps even future artificial moral agents.

Finally, a third vision of machine ethics revolves around a more humanistic pursuit: how to ethically optimize the design of an *amoral* artificial agent? In this final sense, the purview of machine ethics bleeds into the larger institutional level of the state, and in a certain sense, the (global) society. In a theoretical vein, machine ethics is then akin to an exercise in political and legal philosophy or policy building; in an empirical vein, machine ethics resembles descriptive ethics, moral psychology, or computational social choice theory. Here then, a machine ethicist is someone who either a) makes normative claims about how the design and behavior of artificial agents ought to impact human lives and institutions, or b) makes descriptive claims about the observed impact of this behavior, or the expressed or revealed preferences people have towards this behavior.

Since our line of inquiry concerns current and proximate forms of artificial agent, all three senses of machine ethics necessarily apply—either directly or indirectly—to the general enterprise of *narrow* artificial intelligence: to design artificial agents which intelligently act towards a specific goal or carry out a certain activity in an *Umwelt,* or specific social context. Machine ethics

[2] As Steve Torrance describes this sense of machine ethics "There are many…conceptual questions to be addressed here, and clearly the more philosophical inquiries within [machine ethics] overlap considerably with discussion in mainstream moral philosophy. 'Philosophical [machine ethics]' also incorporates even more speculative issues—including whether the arrival of ever more intelligent, autonomous agents, as may be anticipated in future developments of AI, could lead us to have to recast ethical thinking as such, perhaps so that it is less exclusive human-oriented…and possibly dominating or even replacing humanity"(2011, 117).

then provides three distinct evaluative focal points which together constrain the practice of artificial intelligence: the perspective of what an engineer should and should not do, the perspective of how a machine should and should not behave, and the perspective of how a machine—its behavior or its design—should and should not impact human lives and institutions. Taken this way, what we might call the *hard* question of machine ethics is then:

> **how to ethically optimize the successful design of a machine that acts like a moral agent in a specific Umwelt?**

Whatever our answer, *artificial morality* is the agent program which moves this suitably constrained and optimized artificial moral agent from perception to action in its *Umwelt*. It is an account of an *artificial mind* whose structure solves the hard problem of machine ethics, and whose behavioral output endorses the normative standards which are derived from the three perspectives of machine ethics. To achieve this, artificial morality must have three parts. First, a *criterion of rightness* which provides action-guiding recommendations to the agent in its action selection, e.g., an action is right if, and only if, and because it 'minimizes casualties', or 'generates the highest expected payoff for the least well-off person'. Second, a *decision procedure* which is itself comprised of a) a *theoretical* component that provides the reasoning structure which best achieves this criterion of rightness, i.e. 'evaluate the expected harm inflicted on each individual for each alternative action, and choose that action which minimizes net expected casualties' and b) a *computational* component which is able to accommodate this reasoning structure, i.e. a top-down expert system, a partially observable markovian decision process. When an account of artificial morality can be formalized to this extent, it provides a real-life instantiation of explicit ethical agency.

Thus, taken all together, a maximally satisfactory account of artificial morality—or in another way, its victory conditions—leave us to ponder what we might slyly call the *diamond* question of machine ethics:

> **Which criterion of rightness, theoretical decision procedure and computational decision procedure together provide an agent program that successfully moves an**

ethically optimized machine that acts like a moral agent from perception to action in a specific *Umwelt*?

Typically, a response to the diamond question of machine ethics is beyond the scope of any single machine ethicist. In other words, providing an exhaustive answer requires *interdisciplinary collaboration*, and lots of it. This is because, at the very least, the diamond question requires significant computational, moral, and empirical expertise—but likely also requires adjudication from legal and policy experts, analysis from data scientists, and stakeholder consultation and approval—to answer. In brief, it often takes a village to answer the diamond problem of machine ethics, and in light of this, it should not be all that surprising that many machine ethicists can be seen to lack the physical, financial or motivational resources necessary to contemplate it in its entirety.

Unfortunately, if machine ethicists are unwilling or unable to examine this question from all of its faces, they might unwittingly end up answering one-sided questions. Two such one-sided examples, which are in considerable tension with one another, relate to what we can call the *maximalist* and the *minimalist* question in machine ethics. The maximalist, in adopting a philosophical stance on machine ethics, asks himself how to create the 'right' or 'ideal' artificial moral agent, independently of whether a) the behavior of this agent is amenable to its *Umwelt* or aligns with its purpose-oriented ontology, or b) whether the behavior of this agent is acceptable to the individuals it affects. The minimalist, in adopting a comparatively descriptive or democratic stance on machine ethics, asks himself what type of behavior is expected or preferred by society, regardless of whether this behavior is morally optimal or acceptable.

If we find this distinction to be somewhat shallow, for instance in its potential reduction to the possibility of moral realism in machine ethics, we are not only mistaken, but are also liable to begin asking another one-sided question, one which arises when we neglect the composition of artificial morality. Indeed we will likely assume, either tacitly or explicitly, that what is commonly called the *simple thesis*[3] is true: that the question of *how* to build artificial morality is conceptually separate from the question of which values, principles, norms or dispositions are appropriate or

[3] Gabriel, 2020.

'right' to implement. In simpler terms, that the mind of the artificial moral agent need not be affected by the types of moral responsiveness it is designed to display.

To these relatively major oversights, we can add a few more ancillary hang-ups which may affect our capacity to apprehend the diamond question of machine ethics. First, there is the question of what we will come to call the *place* of artificial morality, which addresses the intersection between artificial morality and artificial intelligence, and asks: under what circumstances does an artificial agent require the use of its artificial morality? Or, how *often* should an artificial moral agent act as a moral agent in its environment, continuously, or only when activated by the presence of certain features or action options? Answers among machine ethicists vary widely, and each answer provides a separate conception of what can count as an ethically salient context, which will in turn affect the types of criterions of rightness which seem appropriate for implementation.

Next, there is the troubling question of 'how much should a machine know?', which sits at the confusing intersection between machine ethics as an exercise in moral philosophy and policy building, and the design of artificial morality. Here, we ask: what types of morally relevant features are necessary for a given decision procedure or criterion of rightness, but are nevertheless forbidden by either ethical design concerns such as privacy or autonomy, or ethico-legal concepts such as human or civic rights? This question points to the trade-offs between what moral *behavior* and moral *design* each require of an artificial moral agent, where the former is often severely limited by the latter.

Finally, there is the important practical question of the *adoptability* of a machine, which is in some ways all encompassing. Specifically, this question points to the role and importance of the user, and how his needs are best addressed. In one way, the question pertains to whether some form of morally admirable partiality is required or permissible in AMAs, and if so, what does this mean for the design of artificial morality? In another way, we must ask whether user-centric behavior, or the participation of the user in the design of this behavior, is an expectation that society holds in reference to a specific type of AMA, and is thus either a) directly morally relevant, or b)

indirectly relevant so as to secure the moral benefits that a high level of technological adoption could afford to society[4].

This thesis does not intend to provide a resounding answer to the diamond problem of machine ethics. Firstly, its principal focus is the implementation of ethical decision procedures in artificial moral agents, or more directly, the design of artificial morality in real-world robots. In this sense, it is less concerned with the applied question of the proper deontology of engineers and AI practitioners, and the types of meaningful control, oversight, or legal responsibility they could hold in relation to the artefacts they design. Secondly, given that this thesis is written from the perspective of moral philosophy—rather than social psychology or computer engineering, for instance—the question of how particular moral theories and concepts can apply to the design of artificial morality—rather than the ethical impact or computational ramifications of these machines—will serve as the principal focal point of our analysis.

Accordingly, part II attempts to answer a question which broaches some, but not all of the facets of machine ethics: what type of artificial morality (and by extension, moral theory or reasoning structure) is both technically feasible and societally (or institutionally) acceptable for implementation in an AMA operating in a specific *Umwelt*? Its purview thus broaching the second and third sense of machine ethics, the design of artificial morality, and the computational limitations of narrow AI. In this way, our approach loosely coincides with the v*alue sensitive design* methodology popular in machine ethics[5], but pays particular attention to the conceptual phase of this process.

[4] Another aspect of this problem, which we will spend comparatively little time on, is the question of whether the designer himself has a responsibility to include or delegate to the user in the design of artificial morality, and how this affects questions of responsibility, accountability, or legal liability for AMAs.

[5] Value Sensitive Design is a "…theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman, Kahn & Borning, 2008, 71). It is comprised of (1) a *conceptual phase* where values are identified and elucidated, (2) an *empirical phase* where the apprehension of these values and their trade-offs are bolstered by empirical research into the *Umwelt* of the imagined artefact, and finally (3) a *technical phase* where computational methods and mechanisms are evaluated in terms of their alignment and espousal of these values.

However, the choice of this particular line of inquiry will prove somewhat problematic, since operating across multiple senses of machine ethics, it must pay particular attention to the tensions between them. *Nolens volens*, this requires us to address many of these aforementioned one-sided questions, despite their relying on mutually incompatible assumptions, concepts, vocabulary, and research aims. In other words, the tensions between the different meanings of machine ethics, while of incredible conceptual value for our purposes, nevertheless run the risk of sending us into a dubious 'garden of forking interdisciplinary paths', thus detracting from the clarity of perspective we will need to adequately answer our question. For this reason, we will need to establish a clear and general argumentative path to pursue, with which these one-sided questions will interact at multiple points.

To this end, our path pursues a type of eliminative argument which attempts to identify the vacant space left over for artificial morality once two types of constraints (or conditions) are adequately understood: *technical* constraints, pertaining to what is actually achievable in terms of the computational structure of narrow AI, and *acceptability* constraints, which pertain to the institutional, societal and individual acceptance of both the design and output behavior of an AMA. In this sense, we assume that the place of moral theory and any recommendations it can provide should be maximally conducive to the limitations that these constraints set forth, or more specifically, that artificial morality understood as an exercise in moral philosophy should not *revise* upon the normative standards provided by acceptability constraints, nor should it ignore any technical constraints which point to its impossibility or intractability in narrow AI. More bluntly then, we assume that what matters morally in AMA behavior is not logically independent from what is technically feasible and what is societally acceptable.

With our position clear, part II is laid out in the following way: in chapter four, we explore the types of technical constraints which can be seen to apply to the design of artificial morality. This requires us to address what are likely the three standard computational frameworks for artificial morality: top-down, bottom-up, and hybrid systems. In so doing, we will broach the one-sided questions of the place of artificial morality, and the viability of the simple thesis. In chapter 5, we examine three different senses of the concept of acceptability: what we will call moral preference, adoptability, and institutional viability. Here, we will address the interdisciplinary

question of how much a machine should know. Finally, in chapter 6, we will question how moral theory can accommodate these constraints, and provide our own conceptual model which attempts to do so.

Importantly, given our preoccupation with machine ethics as an exercise in moral philosophy, all three chapters will contend with what is likely the most popular stance for a moral philosopher to adopt: the general problem of *moral maximalism*, or the idea that artificial morality must emulate the 'right' or 'best' moral theory, regardless of technical, but especially acceptability constraints. In chapter 4, we touch upon this problem only in a light way, through our exploration of the critiques of bottom-up approaches, and the ideal place of artificial morality in an AMA's agent program. In chapter 5, the influence of moral maximalism will be greater, since what moral maximalism appears to require of AMAs is often in direct conflict with what people can be seen to want, expect or prefer in the moral behavior of an AMA. Finally, chapter 6 addresses moral maximalism directly, using it as a starting point from which to derive our acceptability and technically constrained approach, which we will call the Ethical Valence Theory.

CHAPTER

4

Technical Constraints & the Structure of Artificial Morality

Our point of departure, and the subject of this chapter, will be to investigate the basic structure of artificial morality as it is portrayed in the machine ethics literature. Specifically, we will be paying attention to the *computational structure* or *decision procedure* of artificial morality, in an attempt to decipher the types of *technical limitations* that the design of artificial morality faces. We will accomplish this by submitting the three most common accounts of artificial morality to critical analysis: the so-called 'bottom-up', 'top-down', and 'hybrid' approaches[1]. Superficially, these categories are distinguished in terms of the decision procedure which they support, which is to say, each approach to artificial morality denotes its own type of *agent program*, or method by which ethical behavior is produced in AMAs. Implicitly, however, these approaches often betray much more substantive claims about artificial morality, including the appropriate *source* of moral content, and the possibility of moral realism, universalism, and value pluralism. Indeed, once these titles are stripped away, a great tension remains between what we have called maximalism and minimalism:, those who see artificial morality as an extension of the field of ethical theory, often

[1] Allen, Smit & Wallach, 2005: Allen, Varner & Zinser, 2000: Wallach & Allen, 2005: 2008.

understood as tracking good and bad, right and wrong *sub specie aeternitatis*; and on the other hand, those who view artificial morality as a question of public acceptability, and thus tracking user expectations, societal preference, and the 'moral wisdom of the crowd'.

The chapter begins with a brief exploration of the engineer's concept of these approaches—as specific agent programs—calling upon distinctions made in previous chapters. The second section addresses the philosophical conception of these terms, including some of the basic assumptions and shortcomings endemic to them. Generally, our exploration of these approaches will yield a number of structural elements which we will hold central to our discussion of artificial morality, and reveal certain insights into the *nature* of artificial morality which will be explored in subsequent chapters. In section III, we delve into a discussion of the various types of technical constraints that affect the design of top-down artificial morality. We begin by addressing the basic structure of many normative theories in section 3.1, and with these concepts in tow, address two roadblocks endemic to the implementation of these theories: what we will call the problem of *constituency* and 'pushless morality' in section 3.2, and the *place* of artificial morality in section 3.3.

1. *The Engineer's Concept of Top-Down, Bottom-Up and Hybrid Approaches*

Loosely put, an agent program is the method by which an artificial agent moves from its perception of its environment to action within that environment[2]. In this sense, top-down, bottom-up, and hybrid approaches to artificial morality all specify different methods by which an artificial agent moves from perception of its environment, to *moral action* within that environment. Superficially then, we might view these approaches as generating fundamentally different types of artificial moral agents, but this would be too hasty. In effect, while each type of approach specifies a specific agent program, in the larger architecture of the machine itself, these approaches may be combined to achieve efficient action in real-world environments. In order to understand how this happens, it should be important to separate two different senses of these terms: the engineering

[2] Russel & Norvig, 2013, 44.

sense, which strictly addresses the agent program, and the more 'philosophical' sense, which addresses the agent program's link to moral theory and its source of moral content.

Accordingly, a top-down approach in the engineering sense is simply an *expert system*[3], wherein a human programmer constructs an explicit and exhaustive list of tasks and subtasks, which can be "…directly implemented and hierarchically arranged to obtain a desired outcome"[4]. This approach aligns with what we have called, in chapter II, *autonomy as provision*, wherein it is the human programmer's task to discern and predict the entirety of the possible changes in the artificial agent's *Umwelt*, and to construct a 'theory' which tracks ideal behavior across these changes. Unsurprisingly perhaps, bottom-up approaches, in the engineering sense of the term, track what we have called *autonomy as independence*, wherein the robot itself learns ideal behavior in its *Umwelt*, thus extending past the a priori knowledge provided by the programmer. This can be accomplished through various computational techniques, such as forms of reinforcement learning, or machine learning and neural network approaches[5].

While the principal difference between these types of approaches is their dependence on *theory*—where top-down approaches necessarily involve an *a priori* theory of action, and bottom-up approaches only yield indirect or *a posteriori* theories of action, and thus typically learn *a*theoretically about their environment—the bottom-up approaches can be further distinguished by the way in which they learn. In reinforcement learning, the robot typically 'discovers' an optimal action rule through trial and error, reinforcing those actions which yielded a positive reward across the experience of the agent, and dispensing with those which yielded a negative reward, the rewards themselves being relative to the performance measure, or objective function, which defines the goal of the system[6].

[3] Russel & Norvig, 2013 : Dignum, 2018.
[4] Allen, Smit & Wallach, 2005, 2.
[5] Ibid.: Dignum, 2019: Wallach & Allen, 2008.
[6] "The agent's sole objective is to maximize the total reward it receives over the long run. The reward signal thus defines what are the good and bad events for the agent. In a biological system, we might think of rewards as analogous to the experiences of pleasure or pain. Sutton & Barto, 2018, 5: Peters et al., 2013: Evans et al., 2018.

In this sense, if a robot is designed to cross a room in order to retrieve an apple, movements that take the robot away from the achievement of its performance measure (to retrieve the apple) are rewarded negatively, while those that move it closer are rewarded positively. For simplicity, we might call this type of bottom-up approach a *reinforcement* approach. Another type of bottom-up approach, however, does not root the agent's learning in its own experiences, but rather, discovers the optimal action rule through the observation of the behavior of others. In this second sense, popular in machine learning approaches, the robot develops a statistical model of optimal action by pouring over vast training sets of data (images, writing samples, driving behavior etc.) in an effort to identify recurring features and correlations, thereby finding the optimal action or 'answer' according to the samples given[7]. This can occur in a *supervised* sense, by using a labelled data set which allows human agents to track the agent's 'performance', or via *unsupervised* methods, where the agent attempts to identify patterns in un-labelled data. Generally, under these *mimicry* approaches, the robot's behavior is dependent on the implicit features and correlations embedded within its training sets, rather than as a direct result of its own experience.

Thus, from a strictly engineering-oriented perspective, these approaches yield three different types of agent program, denoting either a theoretic approach in the case of top-down styles, or a reinforcement or mimicry approach in bottom-up models. Accordingly, in complex artificial agents such as autonomous vehicles, it is quite common that multiple approaches figure in the overall architecture of the machine. For instance, tactical planning may be accomplished by a top-down approach, however perception of the vehicle's environment and its object classification may be accomplished by bottom-up means. In this engineering sense then, most current artificial agents are 'hybrids', combining aspects of both top-down and bottom-up programming styles.

2. *The Philosopher's Concept of Top-Down, Bottom-Up, and Hybrid Approaches*

While the engineering sense of these terms is somewhat generally agreed upon, the more philosophical interpretation of top-down, bottom-up, and hybrid approaches to artificial morality

[7] Kearns & Roth, 2019.

is decidedly more vague. In effect, owing perhaps to the relative novelty of machine ethics and perhaps to its pluridisciplinarity, these terms are often used to denote the *style* of a particular conception of ethical behavior in AMAs, instead of serving as formal definitions, standards or benchmarks[8].

2.1 *The Philosophical Concept of Top-Down Approaches*

Taking each approach in turn then, in its original elucidation, the term 'top-down' denotes "…any approach that takes the antecedently specified ethical theory and analyses its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory[9]". This seems to suggest that a top-down approach to artificial morality is akin to the realization of a successful computational model of a given moral theory, yielding for instance, specifically 'Kantian', 'Utilitarian', or 'Smithian' agent programs. This is to say that models such as these aim to implement the *precise* criterion of rightness and theoretical decision procedure of a given moral theory into a top-down computational decision procedure. While there is a wealth of (early) literature which attempts to do just this[10], more recent work has moved away from staunch implementations of expressly *moral* theories, and towards looser accounts of rule-governed behavior[11].

One important reason for this shift is surely related to the stringency and inflexibility of what Bernard Williams calls the *constituency* of a moral theory. Just as the standard view of moral agency makes serious demands on the ontological characteristics of machines (requiring consciousness, free will, or substantive autonomy), so too do standard moral theories; requiring specific motivations, inclinations, faculties or informational constraints which grant specific entities entry into the 'club' of constituents acknowledged by a given moral theory. According to Williams, for instance, the constituency of contractualist accounts of moral theory is limited to

[8] For an entertaining commentary on this problem, see Marija Slavkovik's interview in the podcast series, *Machine Ethics*, available at https://www.youtube.com/watch?v=VeHKGkTMpJY.
[9] Allen, Smit & Wallach, 2005, 2.
[10] Anderson & Anderson, 2014: 2015: Bringsjord & Taylor, 2012: Powers, 2006: 2013: Gips, 1995.
[11] Aldewereld et al., 2010: Gerdes & Thornton, 2015: de Sio, 2017.

those agents who can conceivably justify their actions to others, and to those agents who can receive these justifications. This amounts to a rather exclusive club of constituents, one which ostensibly denies entry to small children, animals, or the mentally handicapped[12]. On the other hand, utilitarian accounts of moral theory lead to a very broad club of constituents, allowing entry to all entities for which things can conceivably go better or worse (a capacity for welfare), or who can experience pleasure or pain[13]. Leaving the special case of utilitarianism aside[14], the conditions of constituency of most moral theories tend to approximate the conditions of moral agency, and in this sense, it should not be surprising that (current) artificial moral agents fail to gain entry into most, if not *all* clubs, for lack of requisite ontological features[15]. This alludes to the idea that no standard moral theory can be purely 'implementable' in the way that top-down approaches would seem to require.

Nevertheless, the problem of constituency, much like the problem of moral agency, finds its most conscientious objectors within the ranks of moral philosophers. Indeed, barring this specific pocket of machine ethics, the rest of the field can be seen to have taken something of an *interpretive* turn in their approach to top-down programming. This subsequent approach looks less at the stringent application of a pre-existing moral theory, and more to the 'heterogenous grab bag' of applicable rules, constraints, or maxims that could be seen to apply to a specific type of AMA, operating in a specific environment. While these approaches may look to moral theories for inspiration and guidance as to a) the types of morally relevant features which might be present in a given environment, or b) the appropriate criterion of rightness for an AMA's behavior, they necessarily involve some degree of interpretation or adaptation to the purpose-oriented ontology and *Umwelt* of the machine in question.

Across the literature, the degree of interpretation can vary widely. To this end, a relatively tight interpretive account of a top-down approach to artificial morality can be seen in Derek

[12] Williams, 2011, 84. Williams goes on to specify that this exclusivity can be softened by the concern constituents may have for these entities, where the true constituents then act as 'trustees' for this broader class in the representation of their interests.
[13] Ibid, p. 85.
[14] Talbot, Jenkins & Purves, 2017.
[15] Purves, Jenkins & Strawser, 2015: Gabriel, 2020.

Leben's article, '*A Rawlsian Algorithm for Autonomous Vehicles'*, where he presents a model of artificial morality based upon the contractarian tradition of John Rawls[16]. In his plainest terms, he describes this approach thusly:

> The basic idea of this Rawlsian algorithm will be to gather the vehicle's estimation of probability of survival for each player in each action, then calculate which action each player would agree to if he or she were in an original bargaining position of fairness. I will employ Rawls' assumption that the Maximin procedure is what self-interested agents would use from an original position[17].

Plainly, the fact that the autonomous vehicle itself is not a 'self-interested agent' does not seem to irk Leben, nor is he bothered by the fact that Rawls himself certainly did not devise the concept of the original position for use within the context of autonomous vehicle decision-making. Instead, Leben adopts an interpretive approach to Rawls' theory, retaining only a) the probability of survival as the morally relevant feature to consider—a 'best-fit' interpretation of what Rawls' concept of a 'primary good' likely entails within the context of autonomous vehicles, and b) the maximin procedure as the appropriate criterion of rightness, advising the vehicle to choose that action with the highest payoff for the person with the lowest probability of survival, thereby 'maximizing' the minimum share[18]. It is important to note that even if Leben's contribution 'tightly' tracks Rawlsian theory as it might apply to autonomous vehicles, he must nevertheless fundamentally modify the Rawlsian conception of the structure of morality to suit his purposes. This claim is most clearly substantiated by the fact that Leben provides a decidedly consequentialist and monistic account of Rawls, which seems to undercut the plurality of primary goods inherent to Rawls' original theory, and the more deontological origins of his thought. Furthermore, Leben assumes that the principles chosen behind a veil of ignorance for general application in a *society* correspond to the principles that might be similarly chosen for application within the context of *autonomous vehicles*; a substantive claim which is not often made in other

[16] Rawls, 2001.
[17] Leben, 2017, 2.
[18] Leben's original contribution, in reality, advocated the use of both maximin and its lexical sister, leximin, as a further adaptation of Rawls' theory to autonomous vehicle decision-making. Leben's contribution (and interpretation) is hotly contested in (Keeling, 2018).

applications of Rawlsian theory in AI[19]. In actuality then, he leaves us with an autonomous vehicle whose ethical behavior is *inspired* by Rawlsian theory—and specifically the maximin principle—without the need to build the decisional structure necessary to emulate Rawls' theory, or the computational faculties necessary to grant entry into Rawls' club of constituents[20].

To this end, more engineering-oriented proposals of top-down approaches to artificial morality are typically even more brazen in their interpretation of moral theory, and are relatively agnostic regarding the appropriate morally relevant features and ideal criterion of rightness. Perhaps in an effort to subdue some of the long-standing tension between 'rival' classes of moral theory such as deontology and consequentialism[21], some authors have attempted to incorporate multiple theories in their computational models. A shining example of this type of approach can be seen in Sarah Thornton and Christian Gerdes' work on autonomous vehicles. Their proposal advocates for the interpretation of a deontological doctrine—provisionally, Asimov's three laws of robotics—as constraints on the vehicle's movement, combined with a consequentialist-type 'cost' function which weighs contextually relevant values such as mobility, comfort, and adherence to traffic codes[22]. This decision is as much a result of the technical challenges of real-world implementation as it is a will to account for the different ways in which morally relevant features figure into what morality requires of autonomous vehicles, recognizing for instance, the intuitive priority of the protection of human life over strict adherence to traffic codes[23]. Thus, the most appealing aspects of deontological theory (strict adherence to categorical rules) are combined with the most appealing features of a consequentialist-type theory (maximizing whichever values are seen to be relevant to the ideal performance of an autonomous vehicle), all with the aim of ensuring efficient and implementable solutions to autonomous vehicle decision-making.

[19] Indeed, the link between context and the variability of principles selected is typically heralded as a *virtue* of Rawlsian approaches (Gabriel, 2020: Cohen & Sabel, 2006).
[20] To this end, John Sullins (2006) makes a convincing argument that some types of AMAs may indeed gain entry into Rawls' club of constituents, given the theory's heavy dependence on rationality as the key to bargaining in the original position. Whether or not (Leben's) autonomous vehicle satisfies these conditions remains unclear.
[21] Indeed multiple authors have pointed to the drawbacks of implementing a single type of moral theory into a top-down approach to artificial morality: Gerdes & Thornton, 2015: Lin, Bekey & Abney, 2008: Goodall, 2014: Wallach & Allen, 2008: Lin, 2014a.
[22] Gerdes & Thornton, 2015.
[23] Ibid., p. 97.

Thus, while the interpretive turn in top-down approaches to artificial morality has granted designers a fair amount of leeway as to the adaptation, configuration and contextualization of various moral theories, these approaches are nevertheless committed to the implementation of some type of pre-existing theory of moral action or behavior which is designed and implemented by human agents. To this end, Virginia Dignum has provided what is likely the most straightforward set of conditions for a top-down approach to artificial morality—while remaining agnostic about the ideal moral theory to implement. Nevertheless, her account betrays some important implicit assumptions.

For Dignum, there are three conditions which these approaches should meet. First, the software of the agent must "…possess representation languages rich enough to link domain knowledge and agent actions to the values and norms identified…"[24]. This condition coincides with the capacity for an AMA to perceive the (moral) 'marks of significance' of its *Umwelt*, and ensures that the agent is able to register whichever morally relevant features are prescribed by the implemented theory[25]. In this sense, a utilitarian top-down approach would, in a broad sense, only identify 'welfare' or 'utility' as a morally relevant feature, and whichever marks of significance would contribute to the calculation of this feature. This implies a relationship of dependency between a) the moral theory chosen for implementation, b) the morally relevant features prescribed by the theory, and c) the marks of significance that the AMA must be equipped to recognize. It also suggests, however, that a fair amount of energy within the top-down design process must be dedicated to investigating what can count as a 'mark of significance' for a given moral theory, or put another way, what, given the specific context of implementation, can reasonably be considered to instantiate the morally relevant feature prescribed[26].

This has interesting implications for the choice of moral theory implemented, one reason being that it undercuts the superficially simplistic appeal of monistic moral theories such as

[24] Dignum, 2019, 77.
[25] This process amounts to 'extensionally defining principles' (McLaren, 2003), or constructing 'counts as' relations (Aldewereld et al., 2010) between abstract norms and concrete system or environment states, and is a process we will revisit in greater detail in chapter V.
[26] This process amounts to what we will come to call the 'problem of artificial moral uptake' in chapter V.

utilitarianism, which hold only one feature to be morally relevant[27]. Thus, the interpretation of what should count as the instantiation of a value such as 'welfare' or 'utility' across contexts of implementation will vary widely, and as we shall see in later chapters, may lead designers to consider certain ethically dubious characteristics of human agents—such as age, occupation, or criminal background—to be salient instantiations of this value[28]. Finally, the moral problems relating to the 'significance' of these features in the agent's environment are further aggravated by the technical challenges of their reliable detection and classification[29], as well as their overall burden on the real-time efficiency of the AMA's decision-making. Thus, Dignum's condition may prove more restrictive than it initially appears, since it assumes that the agent program of any top-down approach must be able to perceive, classify and interpret the full range of the marks of significance prescribed by a moral theory, without running into intractable problems of complexity[30].

Second, Dignum holds that any top-down approach to artificial morality must be equipped with "…planning mechanisms appropriate to the practical reasoning prescribed by the theory…"[31]. What Dignum aims for here, is a degree of concordance between the structure of the theoretical decision procedure and the subsequent structure of the computational decision procedure. Yet as we have just seen with Leben's Rawlsian algorithm, this condition, in practice, is not as stringent as Dignum makes it out to be. Indeed, since human agents are the presumed constituents of moral theory, some deviation from the structure of practical reasoning prescribed by the theory is always necessary. As established above, some of this deviation can occur within the interpretation of the moral theory so as to align with design specifications, and the purpose and environment of the agent. Another sort of deviation, however, occurs when designers must 'fill in the blanks' which exist between the structure and prescriptions provided by a theory, a task which is both increasingly necessary, and increasingly difficult, as the AMA encounters instances of moral dilemma.

[27] Kagan, 1989 : 1992.
[28] Bonnefon et al., 2016 : Awad et al., 2019.
[29] Keeling et al., 2019.
[30] As Keith Abney maintains, "…top-down theories require an impossible computational load for robot decision-making, due to the requirements for representing knowledge of the relevant effects of action in the world, the difficulty of estimating the sufficiency of initial information, and knowledge about the psychology of other agents and their causal consequences"(Abney, 2012, 45).
[31] Dignum, 2019, 77.

Incidentally, much of the early work in machine ethics was both cognizant and enthusiastic about this required *over*determination of moral theory, believing that it held the "…potential to revolutionize the philosophical study of ethics"[32]. Whether or not this optimism persists today, these authors were clearly responding to a perennial and often overlooked vacuum in moral philosophy, for which the design of AMAs unfortunately required a solution[33]. In one description of this vacuum, overdetermination is necessary so as to provide action recommendations in situations where a moral theory yields no permissible actions in a given decision-context, and thus the agent faces a moral dilemma[34]. Less technically, this problem is described as a situation where 'rules conflict', and is a standard critique of (particularly deontological) top-down approaches in the machine ethics literature[35]. The root of this problem lies in the mobilization of what Robert Nozick calls 'exceptionless moral principles', which while often heralded for the 'safe' and 'idealistic standards' they supply[36], leave open the possibility that an action will possess a combination of features that make it simultaneously required and impermissible[37]. One escape from such an impasse—and the route taken by Gerdes & Thornton—is to transform these 'constraints' into high (violable) costs, thereby breaking up the log jam between conflicting action recommendations[38]. While this consequentialization of a deontological moral theory[39] is certainly feasible from a computational perspective, it necessarily indicates further divergence from the

[32] Allen, Wallach, Smit, 2005, 1. See also Anderson & Anderson, 2011: Berreby et al., 2015: Ganascia, 2007: Moor, 2006: Pereira & Saptawijaya, 2007.
[33] "Research in machine ethics, which of necessity is concerned with the application to specific domains where machines could function, forces scrutiny of the details involved in actually applying ethical principles to particular real-life cases" (Anderson & Anderson, 2007, 16).
[34] Dietrich & List, 2017. In his work on the structure of normative ethics, Shelley Kagan points out this same problem, maintaining that most foundational moral theories lack a 'tradeoff schedule in complex cases involving conflicting factors', and that "…in fact there is little more than a hand or two waved in the direction of showing that the given foundational theory will yield anything like the list of normative factors that we are independently inclined to accept" (Kagan, 1992, 226).
[35] Abney, 2012: Allen, Smit, Wallach, 2005: Wallach & Allen, 2008: Allen, Varner, Zinser, 2000: Dignum, 2019: Gerdes & Thornton, 2015.
[36] Allen, Varner, Zinser, 2000, 260.
[37] Nozick, 1981, 476.
[38] Gerdes & Thornton, 2015. "From a mathematical perspective, dilemma situations represent cases that are mathematically infeasible. In other words, there is no choice of control inputs that can satisfy all of the constraints placed on the vehicle motion. The more constraints that are layered on the vehicle motion, the greater the possibility of encountering a dilemma situation where some constraint must be violated" (95).
[39] Dietrich & List, 2017.

original moral theory implemented, and in this sense, the resulting 'planning mechanisms' used in the agent program differ greatly from those recommended by the moral theory.

Here again then, we might be inclined to revise the high degree of fidelity to moral theory which Dignum demands towards a more interpretive approach, along the lines of Leben's Rawlsian Algorithm. Then, the elements of practical reasoning which must be honored by a top-down approach minimally entail a) the morally relevant feature(s) prescribed by the theory, and b) the criterion of rightness prescribed by the theory. We will address this question of which structural features of a moral theory survive implementation in detail at the end of this chapter. Here, it is sufficient to maintain that all moral theories which presuppose moral agents require some degree of revision and adaptation if they are to be used in top-down approaches to artificial morality.

Finally, Dignum's third condition for top-down approaches to artificial morality posits the need for "…deliberation capabilities to decide whether the situation is indeed an ethical one"[40]. Among all of her criteria, this condition is the most overlooked and under-standardized in the machine ethics literature. This is because, while this condition is certainly plausible, it makes a number of philosophical and computational assumptions which do not categorically hold across proposals of top-down approaches. The main point of contention goes thusly: by maintaining the need for the 'detection' of an ethical situation, Dignum assumes that there are at least some contexts—within the broader set of contexts contained within the AMA's *Umwelt*—which do not require ethical responsiveness. She assumes that the need for artificial morality is 'triggered' by certain events, the presence of certain entities, or the convalescence of a number of action features among the action choices available to the agent. In other words, Dignum holds the view that ethical responsiveness is not a permanent feature of the behavior of an AMA, but rather, is required only in certain circumstances.

This account of what we might call the *place* of artificial morality is hardly generally agreed upon within the field of machine ethics. Indeed, some proposals for top-down approaches to artificial morality—especially those hailing from moral philosophers—do not make a distinction

[40] Dignum, 2019, 77.

between situations of ethical salience and situations of ethical 'indifference', and rather assume that the artificial morality of the AMA—regardless of the type of approach implemented—will remain constantly active in the agent's deliberation[41]. One reason for this may again be attributed to the presupposition of moral agency in moral thought, a traditional assumption of which some authors struggle to avail themselves when considering the agency of machines. This is to say that within traditional accounts of morality, *if* a given human agent is a constituent of a given moral theory, then he is *always* a constituent, unless his moral agency is destroyed, thwarted or temporarily impaired. Moral agency is then a question of *status*, not of context. It seems, in other words, quite strange to imagine a human agent whose moral nose remains completely dormant, or 'switched off' unless it is 'triggered' by the presence of certain features in a decision context, since this would clearly undercut the *universal* type of agency humans are typically seen to possess. Thus, if one neglects the conceptual ramifications of the *modularity* of artificial moral agents, one likewise neglects to consider the *place* of artificial morality in their practical decision-making.

Another, more incisive reason to assume that ethical responsiveness is a permanent feature of AMA behavior hails from an entirely different conceptual hurdle, one which arises whenever we ask the question 'What can count as an ethically salient context?'. This second question is directly linked to a) what the designer—and subsequently, what the moral theory—holds to be morally relevant features in an AMA's environment, and b) the type of moral responsiveness that artificial morality is meant to ensure. In the literature, there is no explicit consensus or general account of the types of harms, evils or unwanted events that we must seek to avoid via the implementation of artificial morality[42]. Instead, claims about this subject vary wildly, from highly inclusive accounts which hold that "Any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation, where the agent finds itself presented with a moral dilemma where any choice of action (or inaction) can potentially cause harm to other

[41] Leben, 2018: Talbot et al., 2017: Gips, 1995: Abney, 2012: Anderson & Anderson, 2010: Grau, 2006.
[42] "…what exactly is meant by morally salient contexts is unclear. For some researchers this would include contexts such as healthcare, elder care, childcare, sex, and or the military—where life and death decisions are being made on a daily (or hourly) basis…For others, morally salient contexts [are] much broader than a pre-defined space or institution…" (van Wynsberghe & Robbins, 2019, 723).

agents"[43]; all the way to highly restrictive accounts which equate ethically salient contexts with lethal moral dilemmas, or so-called 'trolley cases'[44].

One, *a*theoretical way to address this problem is to question what should count as an act of 'harming' in AMA behavior, and seek a substantive definition of a 'harmed-state' for the human agents present within a given *Umwelt*[45]. Another, theoretical way to address it is by evaluating the action-guiding recommendations of a given moral theory, thereby identifying the expected frequency of required or forbidden actions within a given *Umwelt*. In this sense, it would seem fairly easy to assume that ethical responsiveness is a permanent feature of AMA behavior if a) the designer's substantive definition of harm is relatively broad or permissive, i.e., counting 'disappointment' as a moral harm, or b) if the implemented moral theory is relatively 'demanding' in the special philosophical sense of the term[46], leaving little to no room for the pursuit of the agent's practical goals outside of the requirements of morality. Conversely, it will seem easy, if not necessary, that ethical responsiveness be highly punctual if one's substantive account of harm is highly restrictive (say, by tracking only grave physical or lethal harm to human beings), or if the type of moral theory implemented focuses mainly on specific *moral rules* (such as do not kill, do not deceive) rather than moral *ideals* (promote the general welfare, promote equality)[47].

---

[43] Sheutz, 2016, 517.
[44] Wallach & Allen, 2008: Keeling, 2019.
[45] van Wynsberghe & Robbins, 2019.
[46] The problem of 'demandingness' is a critique typically levelled against utilitarianism (Williams, 2011), since the action-guiding recommendations of the theory leave no room for a) special obligations which could thwart the pursuit of the common good, or b) agent-centered options which could limit the degree of sacrifice an agent must endure in the pursuit of the common good (Kagan, 1991). In simpler terms, a moral theory is demanding if it leaves little room for the pursuit of the agent's interests, commitments or subjective values, or forbids special treatment for friends, family, and the like, requiring instead that the agent forgo these things in the pursuit of some impersonal value, such as the common good. It may initially seem queer to address the problem of demandingness in the context of *artificial* moral agents—who could not reasonably be seen to have special ties or personal commitments—but as we shall later see, this critique is surprisingly potent.
[47] "Since acting on any moral ideal is intentionally acting so as to avoid, prevent, or relieve the suffering of harm by someone protected by the moral system…Unlike the moral rules, people are only encouraged, not required, to follow moral ideals"(Gert, 2004, 23). This distinction, which bears significant connection to the problem of demandingness or negative and positive duties (Belliotti, 1981), points again to the appropriate amount of agentive resources a robot must spend on acts which benefit others, or on acts which when performed by humans, may appear as supererogatory or virtuous. The unique ontology and status of robots—most prominently, their lack of personhood—makes this an interesting question in machine ethics, one which we will broach at multiple points across part II.

Thus it seems not only that the question of *place* is essential to the design of top-down artificial morality in the overall architecture of the AMA, but also that this question cuts two ways: first, *place* refers to the role of a top-down decision-making system in the agent program of an AMA—deciding whether it is a separate module, or an integral and constant part of the agent's movement from perception to action. Second, *place* refers to the frequency of ethically salient contexts within an AMA's *Umwelt*, and in a related sense, the morally relevant features which indicate the presence of such contexts, and their substantive definitions. For Dignum, it would appear that her account of the *place* of artificial morality is somewhat restrictive; and accordingly, she sees the need to introduce top-down programming as a separate decisional module in the architecture of an AMA, one that only 'activates' when the agent detects the presence of certain morally relevant features, or when certain rules governing regular tactical planning cannot be jointly satisfied. This stance is popular in certain, more traditional corners of machine ethics, especially those which address the design of autonomous vehicles[48] and autonomous weapons[49].

To summarize then, the hallmark of a top-down approach to artificial morality can be seen as an extension of the engineering concept of an *expert system*. Top-down approaches, like expert systems, rely on a theory, or set of rules or prescriptions, which act together to ensure ideal behavior in AMAs. The role of the theory implemented is then threefold: first, it provides the prescriptions which will govern the agent's behavior, second, it provides some aspects of the structure of the agent program to be implemented, and thirdly, in the case of *moral* theory, it may indicate its proper place in the overall architecture of the machine. Importantly, given that top-down approaches tightly track the concept of *autonomy as provision*, the major benefit of the implementation of a top down approach is clearly its relative determinism: the moral theory implemented is at the very least *chosen* by a human being, and is typically submitted to subsequent tinkering and adjustments to ensure optimal performance. In this sense, human agents are very much at the helm of the behavior of these types of agents, and are able to decide which types of values it holds relevant, and the types of behavior it will likely display. This fact prompts the common remark that top-down approaches are more reliable, predictable, or generally 'safer' than

[48] Evans et al., 2020: Keeling, 2018. A more inclusive account is argued by (Himmelreich, 2018).
[49] Arkin, 2009.

their bottom-up counterparts[50], and bolsters the likelihood of their alignment with ethical design concerns such as accountability, transparency and responsibility.

2.2 *The Philosophical Concept of Bottom-Up Approaches*

The relatively 'safe' characteristics of a top-down approach to artificial morality cannot be properly understood without looking at their main rival: bottom-up approaches. Regrettably, unlike top-down approaches which—while varying widely in the *types* of theories used for implementation and their implications—nevertheless coalesce around their usage of explicit criterions of rightness and theoretical decision procedures, 'bottom-up' approaches regroup a heterogenous body of approaches in machine ethics. Indeed, similar to the engineering sense of the term, the main characteristic shared by bottom-up approaches is of a negative nature: their *lack* of dependence on pre-existing (moral) theory. Thus, within the more thorough accounts of the taxonomy of bottom-up approaches, there exists a host of irreducibly different characters.

From one corner of the literature, comes the contributions made from the field of what is typically called Artificial Life or ALife. The grounding assumption of these approaches is that a science of sociobiology might give rise to a precise account of the evolutionary origin of ethics[51], and accordingly, genetic algorithms are introduced into simplified computer environments (or artificial communities), the interactions of which eventually yield the emergence of behavioral rules and values which might inform real-world decision-making[52]. Interestingly, this approach may offer support to the more choice-theoretic or game-theoretic corners of top-down approaches to artificial morality[53], since the emergent values of the former may spell vindication of the decisional principles of the latter. From another corner, comes the more reinforcement-oriented accounts of bottom-up morality, those that seek to simulate a 'moral education' in AMAs, through embodied learning[54], 'Quest Ethics', or by generally training the AMA via rewards and

[50] Wallach & Allen, 2008: Dignum, 2019: Abney, 2012: Allen, Varner, Zinser, 2000: Allen, Smit, Wallach, 2005.
[51] Allen, Smit, Wallach, 2005.
[52] Theodorou, Bandt-Law & Bryson, 2019: Salimans et al., 2017: Sutton & Barto, 2018.
[53] Leben, 2018: Sirrot & Armstrong, 2008.
[54] Dennett, 1998b.

punishment, approval and disapproval[55]. One particularly *a*theoretical contribution of note among these latter sorts of approaches comes with the idea of 'inverse reinforcement learning'[56], where the agent extracts a reward function given observed optimal behavior, via an 'apprenticeship'[57] with a human expert, or from large data sets[58].

Given the recent rise in popularity of stochastic or machine learning approaches to artificial intelligence in general however, it is quite natural that a good portion of recent accounts of bottom-up morality propose some type of statistical analysis of a 'morally salient' data set, identifying features, patterns and correlations which together yield a decision procedure for ethical decision-making. In one such approach, proposed by Conitzer et al., moral philosophy and psychology literatures are used "…to identify features of moral dilemmas that are relevant to the moral status of possible actions described in the dilemmas. Human subjects can be asked to make moral judgements about a set of moral dilemmas in order to obtain a labelled data set. Then, we can train classifiers based on this data set and the identified features"[59]. This would constitute a supervised mimicry approach to artificial morality. With a hubris similar to that of top-down supporters, these approaches are seen to potentially "…identify general principles of moral decision-making that humans were not aware of before. These principles can then be used to improve our moral intuitions in general"[60]. Under this particular approach then, it would seem that the resultant algorithms are expected to track—if not discover the structural features of— *common sense morality*, or the pre-theoretical moral beliefs and considered moral judgements that drive the everyday decisions of human agents.

This will to track common sense morality sheds light on an important distinguishing feature which varies widely across bottom-up approaches: their variable espousal of moral realism[61], or in a weaker sense, their relationship with the moral 'wisdom of the crowd'. In effect, since top-

[55] Wallach, Allen, Smit, 2005.
[56] Ng & Russel, 2000: Fisac et al., 2020.
[57] Abbeel & Ng, 2004.
[58] Vasquez et al., 2014.
[59] Conitzer et al., 2017, 4831.
[60] Chaudhuri & Vardi, 2014, 1.
[61] Moral realism is the thesis that morality exists independently of human judgements about which actions are wrong or permissible, or that it is possible that every human being is incorrect about which actions are wrong or impermissible (Leben, 2018, 153-154).

down approaches typically track a given moral theory, we should have no trouble expecting the artificial moral agents endowed with these approaches to behave morally. Indeed, if the implementation is successful, the resulting AMA will *perfectly* emulate the action-guiding recommendations of operative theory, yielding say, a perfect utilitarian AMA[62]. Then, if all the world were populated by similarly perfect utilitarian *human* agents, the expectations of the users interacting with the robot would perfectly align with the robot's behavior. In other words, the judgements which inform the actions of the utilitarian AMA would be praiseworthy[63], valid or *acceptable* to each and every user it comes across.

Yet, what happens when the utilitarian AMA is confronted with a different world, one in which half the population are perfect utilitarians, and the other half are pure Kantians? In this world it would seem, the actions of the utilitarian AMA, despite being morally 'right' by utilitarian standards, will nevertheless be *unacceptable* to the Kantians, such as what might happen when the utilitarian AMA elects to kill a user's ageing Aunt Agatha, thereby ensuring that her impressive fortune goes to charity, rather than to the construction of a feral cat sanctuary, as she had planned[64]. This discrepancy between the moral judgements of a population, often identified as the problems of *moral relativism, pluralism, or (lack of) universalism* in machine ethics[65], leads us naturally to what is likely the principal virtue of bottom-up approaches to artificial morality: where top-down approaches are generally unable to address moral acceptability—since they typically impose one pre-existing moral theory regardless of public moral sentiment and judgement—bottom-up approaches cater directly to it, typically through its *a*theoretical assessment and aggregation into workable decision principles. In simpler terms, the glory of such bottom-up approaches lies in the incorporation of *all* of the morally relevant features and criterions of rightness which could be seen

[62] Gips, 1995.
[63] In the literature, this term is often thrown around as a tacit design goal of artificial moral agents. We will address its implications in chapter VI. Here, we will define it simply as an instance where the decisions of an AMA appear to validate the normative or moral expectations of the affected human users.
[64] The case of 'Aunt Agatha' and the perfect utilitarian AMA is generously explored in Talbot et al., 2017, and we will also give it a generous amount of attention across part II.
[65] Wallach & Allen, 2008: Rawad, 2019: Bonnefon et al., 2016: Abney, 2012: Dignum, 2019: Allen, Smit, Wallach, 2005: Noothigatu et al., 2018: Beavers, 2011.

to motivate the moral intuitions of a given population, yielding an artificial morality which in many senses *mimics* this population's collective moral nose[66].

Here again, the method by which these bottom-up approaches pass from the collection of considered moral judgements, sentiments or preferences to the generation of a workable decision procedure can vary. However, in recent years, empirical studies in the form of *serious games,* such as MIT's Moral Machine Experiment, have gained in popularity[67]. Under this latter approach, the moral intuitions of test subjects are subjected to trolley-type dilemmas, where the subject must choose who to sacrifice or spare in the event of an unavoidable collision with an autonomous vehicle. Then, with this data set of moral preferences in tow, the AMA will "…learn a model of societal preferences, and, when faced with a specific ethical dilemma at run time, efficiently aggregate those preferences to identify a desirable choice"[68], sometimes aided by the mobilization of a specific theory of voting rules[69]. In this sense, many current bottom-up approaches take heavy inspiration from theories of computational social choice[70].

Importantly then, while the philosophical sense of bottom-up approaches can be seen to align with its engineering sense, in so far as a) the morally relevant features and criterions of rightness which guide the AMA's decision-making are discovered *a*theoretically[71], and b) the process by which this occurs aligns with either a reinforcement or mimicry approach in the engineering sense of the term; there is, at a deeper level, a sense for which 'bottom-up' is synonymous with 'grassroots' in the question of the *source* of moral content. Put another way, one of the hallmarks of decidedly bottom-up approaches to artificial morality is a feeling of rebellion against what Henry Sidgwick calls, the 'science of conduct', or the rational, stringent and internally

[66] This inability of many moral theories to track the full breadth—or in some cases, even a significant portion of—the considered moral judgements of users and society will later come to form one of our principal critiques of the 'maximalist' approach to artificial morality.
[67] Bonnefon et al., 2016.
[68] Dignum, 2019, 78.
[69] Prasad, 2018. These voting approaches are seen to align with the so-called parliamentary model of moral decision-making (MacAskill, 2016), where individuals assign probabilities to the likelihood that different moral theories are true, and then estimate the choiceworthiness of options on that basis (Gabriel, 2020).
[70] Greene et al., 2016: Gabriel, 2020.
[71] We should be careful not to confuse the use of 'voting theories' in many bottom up approaches with the use of moral theories in top-down approaches, since the former serve a *procedural* role in the *discovery* of the criterion of rightness, while the latter directly provide this criterion.

coherent principles which inform most moral theories. Indeed, bottom-up approaches, at this general level at least, can be seen to be more *democratic* in their quest to understand ideal robotic behavior in morally salient contexts, and are superficially more inclined to 'give the people what they want' than most of their top-down counterparts. What ought to happen when 'what the people want' includes morally onerous content, immoral biases, or morally dubious inclinations, constitutes the principal critique of bottom-up approaches by their top-down counterparts, and will fuel much of the discussion of chapters V and VI.

Nevertheless, even if we find the democratic zeal of bottom-up approaches appealing, they present, much like the stochastic AI systems from which they take inspiration, supplementary ethical problems which must be addressed. Most principally, bottom-up approaches suffer from a lack of transparency, as to a) the types of features the system will identify as morally relevant, and subsequently b) the types of criterions of rightness which will guide the AMA's behavior[72]. In this sense, the behavior of a bottom-up AMA is substantively less predictable than its top-down counterpart, and it may act on principles which are neither easily divined nor easily formalized. Furthermore, this opacity can migrate into the original intentions of the designer, since "…little can be inferred about the intent or conduct of the humans that created or deployed the AI, since even they may not be able to foresee what solutions the AI will reach or what decisions it will make"[73]. In this sense, the ability to produce an accountable, transparent, and responsible machine via bottom-up methods remains particularly challenging. Generally then, these concerns constitute a morality-oriented version of the 'black box' problem endemic to stochastic AI, as explored in the first chapter.

To these general concerns, are added problems which specifically pertain to the 'democratic methods' employed by bottom-up approaches. As Virginia Dignum points out, bottom-up approaches assume both a) that the choice of the crowd is equatable to a 'system of ethics', and b) that a sufficiently large amount of data can indeed be collected from a suitable set of subjects[74]. In the case of (b), the question at stake is whether or not the data collected can be

[72] Dignum, 2019.
[73] Castelvecchi, 2016, 893.
[74] Dignum, 2019, 79.

truly *representative* of the population—and thus truly represent the full scope of common sense morality—or whether it is inherently biased towards the preferences of those individuals who, say, elected to play a serious game[75]. In the case of (a), her claim is slightly more vague, however on one account, it would seem to point to uncertainty concerning whether the responses provided by the test subjects truly constitute pre-theoretical moral intuitions—those things which moral theory ideally tracks, or at least are judged by—rather than simple pre-reflective beliefs or reactions[76]. On another account, it would seem to cast doubt on the internal coherency of the resultant decision principles. In any case, both (a) and (b) serve to undermine the appeal to democracy and public acceptability with which bottom-up approaches secure their supposed supremacy, and in this sense, generate further skepticism as to the desirability of any approach which advocates for the discovery of ideal moral behavior through such independent means.

2.3 *The Philosophical Concept of Hybrid Approaches*

Given the relative heterogeneity of the methods of bottom-up approaches to artificial morality, and given the relative heterogeneity in structure, features, and theories of its top-down counterpart, it would seem to follow that the concept of a hybrid approach—one that combines aspects of both top-down and bottom-up approaches—would be equally disparate. Nevertheless, given the drawbacks of both top-down and bottom-up approaches to artificial morality, hybrid approaches are often seen to be relatively desirable, and are for this reason somewhat common[77]. The reasoning behind this desirability can be seen in the original elucidation of the concept of a hybrid:

> Top down principles represent broad controls, while values that emerge through the bottom-up development of a system can be understood as causal determinants of a system's behavior…Top down approaches emphasize the importance of explicit ethical concerns that arise from outside of the entity, while bottom up approaches

[75] Indeed, in regard to the MIT Moral Machine Experiment, a common—yet understandably unpublished—critique of the results is to hold that they are not sufficiently robust or representative, catering instead to the 'young, white, male gamers' who found themselves drawn to the study. A more elegant version of this same claim is given by Abby Jacques in (Jacques, 2018).

[76] This point can be attributed to Marija Slavkovik, herself a prominent figure in the computational social choice corner of bottom-up approaches to artificial morality.

[77] Conitzer et al., 2017: Arkin, 2008: Bringsjord et al., 2016: Verdiesen, Dignum & Van Den Hoven, 2018.

> are directed more at the cultivation of implicit values that arise from within the entity[78].

This particular definition provides a more engineering-oriented account of a hybrid, since it seems to suggest a combination of adherence to theory, mimicry and reinforcement learning. Metaphorically, a hybrid then appears a bit like Rousseau's *Emile* with stricter parents, wherein the robot is encouraged to act autonomously, drawing its own conclusions and value judgements, so long as this pursuit does not entail the violation of a set of basic moral rules laid down by the programmer. To this end, a straightforward (non-moral) example of this approach can be found in MIT, IBM and Google's 'neuro-symbolic concept learner'[79], which learns to reason about physical phenomena using different colored blocks and shapes, in ways which simulate early childhood education. In detail, the learner uses "…a neural network to recognize the colors, shapes and materials of the objects, and a [top-down] system to understand the physics of their movements and the causal relationships between them"[80]. Initial reports suggest that it far outperforms both top-down and bottom-up systems in testing[81], overcoming the limitations of both approaches.

Thus, within this account of a hybrid approach to artificial morality, it is the structural features of the agent program which are at issue. The top-down theory provides a minimal account of ideal action for the AMA, generating rules that cannot be broken, but much of the complexity of the agent's environment is addressed by bottom-up means, thus blending the most efficient parts of expert and stochastic systems. However, recalling our claim that there exists one sense of the term 'bottom-up' which pertains to the choice of the *source* of moral content, it would appear that a hybrid approach could also denote a blend between the 'democratic' methods of bottom-up approaches, and the stringent moral theories used in top-down approaches. In this second sense, a 'hybrid' would then be an artificial moral agent which blends the most 'efficient' parts of moral theory and common-sense morality within the morally relevant features and criterions of rightness which guide its agent program. Importantly, an AMA which is hybrid in its source of moral content

[78] Allen, Smit, Wallach, 2005, 153.
[79] Mao et al., 2018.
[80] Hao, 2020.
[81] Knight, 2019.

need not be hybrid in its structure[82]. Rather abstractly, if it is possible to *explicitly* discern a criterion of rightness which reflects the moral wisdom of the crowd—through empirical research, or even certain bottom-up methods of preference aggregation—then it is plausible that a top-down system can be built which abides by *two* criteria of rightness: one reflecting 'democratic' moral sentiment, and the other reflecting moral theory. Still, it may not be obvious why this type of hybrid approach is a desirable alternative to any of the models proposed above, beyond the somewhat trivial idea that this type of hybrid—being a top-down approach—would skirt the concerns for the opacity of bottom-up systems.

Yet, if we examine the question of the *source* of moral content more carefully, we might have a very clear reason to prefer this type of hybrid over other standard approaches: its capacity to broach the divide between *ethical behavior* and *public moral sentiment*. In the literature, these two sources of moral content are often at odds, with the stark defenders of one pointing to the moral blind spots of the other. In our terms, this constitutes what is likely the principal quarrel between the maximalist (expounding normative ethics) and the minimalist (expounding descriptive ethics). The battle surrounding the ideal source(s) of moral content is a clear extension of the larger *AI alignment problem* in the field of (general) artificial intelligence[83], itself a more sophisticated version of what we have called the *argument from increasing automation*. Loosely formulated, the AI alignment problem seeks to uncover those principles, theories or sources of moral content which best align with what humans want from AI, and thus indirectly seeks to discover the substantive content of humanity's *coherent extrapolated volition*: what we would want if "…we knew more, thought faster, were more the people we wished we were, and had grown up farther together"[84]. According to some, the victory conditions of artificial morality coincide with our coherent extrapolated volition, even if this argument is typically made within the context of highly advanced AMAs, Artificial General Intelligence, or Superintelligent agents. Importantly, even in the case of the 'narrow' or 'tool AI'[85] which underpins our concept of artificial moral agency, the designer's choice of *approach* to artificial morality will depend heavily on his

[82] This distinction, for instance, was made explicit in Tolmeijer et al. (2020)'s extensive machine ethics implementation survey.
[83] Or alternatively, the *value alignment problem* (Fisac et al., 2020: Gabriel, 2020: Yudkowsky, 2016).
[84] Yudkowsky, 2004, 6.
[85] Bostrom, 2009.

particular take on the AI alignment problem. This will lead him to endorse a particular type of agent program, and subsequently, a particular structure of artificial morality, including: the ideal *source* of moral content, the ideal *place* of moral responsiveness, the ideal *constituency* of artificial morality, the ideal morally relevant features (and marks of significance), and the ideal criterion of rightness—even if some of these structural features are discovered via methods opaque to him.

What this suggests is that the choice of a particular agent program may commit one to a corresponding substantive view of the scope, role and purpose of artificial morality itself. Or conversely, a particular belief about the nature of artificial morality will often lead to the implementation of a specific type of agent program. This stands in stark contrast to what Iason Gabriel has called the *simple thesis*, according to which "…it is possible to solve the technical problem of AI alignment in such a way that we can 'load' whatever system of principles or values we like later on…sometimes [carrying] with it the further tacit implication that the search for answers to the philosophical questions can be delayed"[86]. In the last section of this chapter, we will explore the cogency of the simple thesis, particularly as it pertains to the place of artificial morality. This will provide us with a number of ways in which technical limitations can affect the design of artificial morality.

3. *Technical Constraints & Artificial Morality*

While the previous sections have explored three types of computational approach (or computational decision procedure) which aim at achieving moral output behavior in AMAs, it should be clear why moral philosophers hold a strong preference for top-down or hybrid approaches: they constitute the only options for which moral theory could ostensibly be applied, or the only options yielding a machine which could act and reason like a Smithian, Kantian or Rawlsian type of moral agent. Thus, artificial morality when understood as an exercise in moral philosophy often presupposes top-down or hybrid approaches. In this sense, our discussion of the technical limitations imposed on artificial morality will also assume a top-down or hybrid approach to its computational decision procedure.

[86] Gabriel, 2020, 3.

With this idea in place then, we can usefully revisit the assumptions made by the *simple thesis*. At first blush, it would then appear that the simple thesis assumes a relationship of *independence* between a top-down computational decision procedure on the one hand, and *any* pairing of a criterion of rightness and a theoretical decision procedure, on the other. This however, seems to thwart Dignum's second condition of a top-down approach, namely that the computational decision procedure implemented must have "…planning mechanisms appropriate to the practical reasoning prescribed by the theory"[87]. Even if we maintain, in line with our previous analysis, that a certain degree of interpretation is always required to avoid the problems of the *constituency* of a given moral theory—a conceptual hard line that points to the impossibility of artificial morality for lack of an AMA's ontological capacities, such as free will, autonomy, sentience, intentionality or the ability to act on reasons etc.—there remains the interesting question of *how much* interpretability is typically required in the design of top-down approaches to artificial morality. In effect, approaching the problem this way, we might be able to identify a) certain types of moral theory which are more challenging to implement from a computational point of view, and perhaps b) how much a suitably interpreted theory realistically resembles the moral theory it seeks to emulate, or how Smithian a 'Smithian' AMA actually is.

However, if we are attempting to 'complicate' the simple thesis in this way, it should first be appropriate to identify the basic structural content of a moral theory, so as to discern where exactly this necessary interpretation takes place. To this end, the following section provides a basic elucidation of the general elements present in most (rationalist) moral theories, and then points to two particular limitations that any moral theory must accommodate if it is to be used in the design of artificial morality.

3.1 *The Structure of a Moral Theory*

Thus far in this chapter, we have mentioned a number of structural features which a moral theory can provide to the design of artificial morality: a *criterion of rightness*, which states the right-making feature of an action (i.e. an action is right if, and only if, it maximizes aggregate

[87] Dignum, 2019, 77.

expected utility)[88], a *decision procedure*, a method which achieves this criterion of rightness (i.e. observe the permutations in general welfare across all possible actions, and select that action which generates the greatest expected net gain in aggregate welfare), and finally *morally relevant features*, which provide the facts that ground the decision procedure and criterion of rightness (i.e. facts about welfare such as health, material wealth, happiness, absence of pain, etc.).

To this trio of concepts, we can add, following Shelley Kagan's analysis in *the Structure of Normative Ethics*[89], three further structural elements: what he calls the *normative* and *foundational* level of a moral theory, and a theory's (primary) evaluative *focal point*. The normative level of a moral theory explains what is normatively relevant to the determination of the moral status of an act, or put another way, the types of value that a given moral theory holds to be morally relevant. Consequentialist theories, for instance, only hold one type of value to be normatively relevant, the overall goodness of results; and in this sense provide a *monistic* account of value. Thus, acting as a pure consequentialist, the only factor I must consider in my determination of the moral status of an act is the overall goodness of its results; other considerations, such as whether or not I am keeping my promises or fulfilling my duties, are all *adiaphoric*, or do not have any moral significance when I determine what I ought to do.

Other theories, however, can hold other types of value, or multiple types of value to be normatively relevant. Indeed, most deontological types of moral theory hold that certain types of general constraints (such as human or natural rights) are inherently valuable and normatively relevant, as well as certain types of special obligations (a mother's obligation to take care of her child, the obligation one has to keep his promises, etc.). These types of theory often provide pluralistic accounts of value, which may or may not count general utility, or the overall goodness of results among them. In this sense, if I am acting as some type of deontological agent, multiple factors may influence my assessment of the right action to perform, including whether I would be breaking any promises, intentionally causing harm, thwarting my obligations to others, and so forth.

[88] Brink, 1986.
[89] Kagan, 1992.

Importantly, beyond the notions of overall goodness of results, and special and general obligations, lies a final type of value—which Kagan calls 'options'—relating to whether the *personal cost* of the performance of an act ought to have normative importance (the value being the personal value of an agent's commitment to his life projects or aims[90]). In this sense, 'options' cover much of the same conceptual territory as the *demandingness* of a moral theory[91], since options, or more specifically, *agent-centered constraints* or a*gent-relative permissions*, serve to moderate the limits imposed *by* morality on the individual, or the limits the individual imposes *on* morality, respectively[92]. To illustrate, a moral theory would certainly be very demanding if it held that I was morally required to donate all of my non-essential income to charity, thereby making the greatest contribution to overall welfare, or promoting the best overall results. Under this theory, I do not have the agent-relative permission to fail to promote the good in situations where it would come at considerable cost or discomfort to me. On the other hand, a moral theory is also quite demanding if it forbids me to, say, twist a criminal's arm and thus cause him pain, even if in doing so, his subsequent confession would lead to the dismantlement of Mexico's largest and most violent drug cartel. Under this theory, I am burdened with the agent-centered constraint of never causing (even minimal forms of) harm, even if doing so would bring about the best overall results.

The foundational level of a moral theory, in turn, provides an account of *why* or *how* certain factors can be normatively relevant. In this sense, a theory is consequentialist if it maintains that whatever the normative relevant factors are, the grounds of their relevance are ultimately rooted in their connection with the overall goodness of results. Thus, certain types of utilitarianism or consequentialism may accommodate general obligations such as human rights at the normative level, since at base, their relevance is explained by the overall goodness the existence of human rights would bring about. Foundational theories can also take on a more procedural bent, such as

---

[90] Nagel, 2012.

[91] In a more vague sense, options also track the conceptual difference between *satisficing* and *maximizing* accounts of morality; where the former counts as morally permissible any act which conforms to a few basic moral constraints—leaving more room for action options—and the latter counts only those actions which lead to the best moral consequences as permissible, leaving comparatively little room. (Bostrom, 2014: Ogien, 2007: Nozick, 1974, 500: Gabriel, 2020).

[92] "…I am not required to significantly sacrifice my interests in order to provide aid to others—even though objectively greater good would result from my doing so. Of course, I am *free* to make such sacrifices if I choose to—and morality encourages me to do so—but these acts are not *required* of me…" (Kagan, 1989, 4).

universalization theories, which view the correct list of normative factors as those which can be appropriately universalized without logical contradiction; or contractarian theories, which hold that the correct list of normative factors is that which could be agreed upon by a group of suitably informed or motivated bargainers. In this sense, Rawlsian moral theory is contractualist at the foundational level, since it views the correct set of normative factors as those which would be chosen by individuals behind a veil of ignorance in the original position.

Finally, Kagan mentions the important notion of an 'evaluative focal point', which points to the *object of moral assessment* used to evaluate the moral status of an act. In this sense, not only the act itself, but also rules, motives, institutions, norms, character traits, or intentions can serve as the (or one of many) 'focal point(s)' of moral assessment. In this sense, if I use a moral theory which is foundationally and normatively consequentialist to decide the moral status of an act, I need not necessarily evaluate the *act* itself. Instead, I might look for a set of general rules which bring about the best results, and perform that act which is permissible or required by these rules (a form of *rule* utilitarianism), or I might try to adopt whichever disposition or character trait might lead to the best results. Very roughly, virtue theories often have dispositions or character traits as evaluative focal points, while deontological theories often make use of rules, motives or intentions. Contractarianism, like rule utilitarianism, only evaluates acts indirectly, and directly evaluates the *rules* which would be chosen by, for instance, hypothetical bargainers.

Putting these notions together, we get the following picture: A criterion of rightness explains what to do with the *normatively* relevant factors of a given moral theory, which are themselves arrived at or tethered to the *foundational* moral theory. What we have called *morally* relevant features in turn, pertain to the facts which ground the normatively relevant factors, or how the values which are normatively relevant are instantiated in an environment. For instance, goodness of results is likely grounded in facts about welfare (health, income, capability, etc.), while special obligations are grounded in facts such as, say, my having promised Peter that I would help him set up his new printer this afternoon. What we have called *marks of significance* then can be described as the phenomenal signature of the morally relevant features in the AMA's environment, i.e. 'smiling' is an indication of happiness, and therefore welfare, in a social robot's *Umwelt*. Evaluative focal points, in turn, point to how the normatively relevant factors, and morally

relevant features of a given moral theory are managed by its decision procedure. For instance, if I am an act utilitarian, my decision procedure would likely be to evaluate the permutations in welfare across action options, and to select that option which maximizes welfare. If I am a *rule* utilitarian, then my decision procedure would be to seek out that list of general rules of conduct which leads to the best overall expected welfare, and then to act on those rules. Options, finally, relate to how much a particular theory demands of me as an agent with value, aims and commitments: a theory can provide certain permissions allowing me to fail to do what is morally required if it would come at considerable cost to me, or it could impose serious restrictions on how I pursue my aims in life, forbidding me for instance, from harming any moral agent in the process.

From this vantage point, we can see that some of the structural elements just presented pose relative difficulties for artificial moral agents, in much the same way that the problem of constituency proves difficult for the design of artificial morality. One such area of difficulty lies in a theory's use of certain evaluative focal points such as intentions or dispositions, since it is not clear whether robots could conceivably possess the ontological attributes necessary to *intend* to act, or be *disposed* to act, at least in the way that moral theory prescribes[93]. Thus, theories which make use of such focal points—such as the doctrine of double effect, or many accounts of virtue ethics—will necessarily require a higher degree of interpretation if they are to be implemented[94].

Another such limitation arises when a moral theory makes use of certain types of foundational moral theory, especially universalization theories such as Kant's categorical imperative[95], or contractarian (or contractualist) theories such as Scanlonian moral theory[96]. Taking Scanlon as an example, since this theory views the correct list of normative factors as those

[93] Winfield et al., 2019: Dennett, 1998a: Tonkens, 2009: McDermott, 2008: Gips, 1995: Grau, 2006: Johnson, 1985.
[94] Indeed, resemblance to virtue ethics is usually ascribed to a model of artificial morality *a posteriori*, rather than as the result of an explicit design goal (cf. Lin et al., 2014). However, one *bottom-up model* which takes virtue ethics as an explicit design goal can be found in the 'moral functionalist' approach of Howard & Muntean (2017). Interpretive versions of the doctrine of double effect can nevertheless be found in (Bonnemains, Saurel & Tessier, 2018: Govindarajulu & Bringsjord, 2017).
[95] In his insightful article about the implementation of Kantian ethics, Thomas M. Powers cites a number of challenges—triviality, asymmetry, excessive specificity, lack of semidecidability, and a lack of priority for maxims—all of which must be overcome before a workable model of artificial morality can be attained (Powers, 2006).
[96] Scanlon, 1998.

which would provide reasons for action that are contextually justifiable to individuals existing in a relation of 'mutual recognition'[97], and uses rules rather than acts as an evaluative focal point, the computational load required by this theory likely borders on intractability, at least in real-time decision-making. This is so since the decision procedure mandated by the theory passes through an examination of the implications a given action principle (or rule) has for every individual affected by the agent's decision, and searches out that principle of action which cannot reasonably be rejected by anyone. In this sense, a tight implementation of Scanlonian contractualism requires the generation of multiple sets of action principles (or criterions of rightness), and requires deliberation *across* these sets to find the optimal criterion of rightness given the context. Accordingly, the morally relevant features (or marks of significance) that must first be detected and assessed by the AMA are likely incredibly vast, as is its need for world knowledge and common-sense reasoning[98]. To this end, similar claims have been made concerning the use of utilitarianism in artificial morality, since the high degree of generalization, and the apparently limitless temporal horizon of the theory both spell intractability unless the parameters of the theory can be adequately bounded[99], and thus interpreted for use in AMAs[100].

Of course, the degree of interpretation a moral theory will require for its use in artificial morality, as well as the general intractability of certain moral theories, will both greatly depend on the specific type of AMA and *Umwelt* in question. As we have seen in previous sections for instance, the use of a Rawlsian theory for autonomous vehicle decision-making—while relatively 'tight' compared to comparable contributions in the literature—nevertheless turns out to be highly interpretive, since at least on Leben's account, the resulting computational decision procedure does

[97] This is to paraphrase his decidedly more popular elucidation of this view: "an act is wrong if its performance under the circumstances would be disallowed by any set of principles that no one could reasonably reject as a basis for informed, unforced, general agreement" (1998, 153).

[98] In chapters VI and VII, we will introduce our own account of artificial morality, the Ethical Valence Theory, which nevertheless relies on some Scanlonian assumptions and structuring. Still, it remains a very *loose* interpretation of Scanlon's theory.

[99] Anderson & Anderson, 2007: Allen, Varner & Zinser, 2000: Gips, 1995: Wallach & Allen, 2008: Klincewicz, 2017.

[100] One interesting tangential problem in the implementation of moral theory is the idea that while this implementation may be technically possible—to varying degrees of interpretation—it is nevertheless forbidden, or self-defeating, according to the moral theory in question (Tonkens, 2009: Nadeau, 2006: Torrence, 2008). Typically however, these arguments gain much of their support from the presumption that full ethical agency must be attained for these theories to be implemented correctly.

not make use of a) the correct focal point (rules not acts), and b) the correct normatively relevant factors—either the two principles of justice (rather than simply the maximin principle), or on a decidedly tight interpretation, whatever principles would be chosen by individuals in the original position behind a veil of ignorance, who together decide on the fundamental principles of a just *autonomous vehicle*. In this sense, the foundational theory, too, seems to be somewhat neglected in Leben's work. Thus, since the *degree* of interpretation can vary widely depending on the context of AMA implementation, and to a certain extent, the *awareness* some moral philosophers exhibit as to the interpretation of moral theory in their work, it is difficult to make very many categorical or general claims about the impossibility of certain moral theories from a technical standpoint. All that we can say is that 'tight' interpretations of a moral theory into a computational decision procedure require the implementation of every structural feature of the moral theory, and the more these structural features include aspects like universalization, generalization, or contextual awareness, or focal points such as dispositions or intentions, the more they will border on intractability[101]. Put another way, the more a moral theory requires specific mechanisms which depart from the basic computational structure of an expert system, the more the simple thesis appears false.

Nevertheless, one claim we will make concerns a certain *division of decisional labor* between the designer (or the moral philosopher) and the AMA in the design of a computational decision procedure which expounds a moral theory. In this sense, even if some aspects of the theory—particularly parts of the decision procedure—are not accomplished by the AMA itself, there is nothing inherently immoral or untoward about the designer (or moral philosopher) accomplishing some of this moral deliberation *a priori*, so long as the way in which *he* himself deliberates corresponds to the decision procedure mandated by the moral theory. Put another way, Leben's Rawlsian AV is not less moral or less valuable than the Truly Rawlsian AV we have just described, rather, the difference is one of *technical* sophistication amongst the ranks of explicit ethical agents. In this sense, Leben's Rawlsian AV is surely less 'autonomous' than the Truly Rawlsian AV, since his vehicle simply uses the maximin principle as a criterion of rightness, rather than engaging in the lengthy process of adopting the moral perspective of the original position and

[101] Abney, 2012: Allen, Smit & Wallach, 2005: Allen, Varner & Zinser, 2000.

generating its own principles of justice. However, this relative deficiency in autonomy certainly does not carry much value, in so far as even the Truly Rawlsian AV still fails to meet the necessary conditions of personhood or moral agency, at least according to the standard view. Put simply then, the interpretation of moral theory in the field of machine ethics, whatever the degree, should only be considered a crime if the 'interpreter' fails to adequately understand the theory he seeks to emulate. Some agent (whether human or artificial) ought reasonably to consider a moral theory from tip to tail (or criterion of rightness to marks of significance) if its implementation can truly count as a computational model of the theory, but it makes no *moral* difference which type of agent accomplishes which part of the deliberative workload.

This being said, even if the possibility of the interpretability of a moral theory is in many ways contingent on the practical purpose and environment of the AMA in question, there are two very general technical limitations which can affect the interpretation of all types of moral theory. The following two sections explore these in detail.

3.2 *There Is No 'I' In Robot*

As we have seen in the first section of this thesis, one flagrant way in which artificial moral agents fail to meet the necessary conditions for moral agency is through their lack of *subjectivity*. They have no passions, inclinations, drives, emotions, commitments or autonomously derived principles, nor is there any meaningful sense for which they have a 'self-interest'. For some machine ethicists, this squashes any hopes of true moral agency in AMAs[102]. For others however, this lack of subjectivity is often seen as a *virtue.* Indeed, under a very hard interpretation of 'tool AI', it is certainly plausible that a lack of many of the qualities of personhood would prove useful in the performance of certain tasks or roles: for instance, a robot soldier that has no sense of self-preservation is certainly more tactically useful than a human soldier who does[103]. Under this rather simple view then, a great deal of what makes an AMA perform 'better than' a human agent in specific environments—and thus what makes these AMAs particularly attractive for research and development—pertains to this lack of self-interest or self-preservation. Thus, at a functional level,

[102] McDermott, 2008.
[103] Arkin, 2009.

a lack of subjectivity is useful if it bolsters technical performance, generating a world in which for instance, robotic healthcare workers never sleep, take breaks, vacations, or go on strike, and where (empty) autonomous vehicles drive themselves off cliffs at the first detection of a squirrel in the road.

Moving closer to the purview of morality, a first type of claim we could make is that the fact that artificial moral agents have no inherent *moral* value[104] is instrumentally valuable for humans. Indeed, we might go so far as to say, along the lines of Joanna Bryson, that this arrangement is *morally good*, in so far as the alternative seems to be building AMAs to which we *do* owe certain forms of moral behavior, and who might nevertheless exist in positions of servitude in human society[105]. But if personhood does not, and perhaps ought not figure in the ontological qualities of an AMA, we are consequently forced to examine how this affects the shape of moral theory, and thus the design of artificial morality. What does morality look like when it passes through the agency of an entity with no moral value, status, claims or rights[106]?

Even from a pre-theoretical perspective, we can make the claim that a 'subjectless' agent disrupts the fabric of any moral theory. This is so since it is certainly plausible that *any* moral theory[107]—or for that matter even the pre-theoretical intuitions that inform common sense morality—all presuppose a human agent as a vessel through which the good and the right are carried out, and importantly, to which certain behaviors are owed from others. Put in simpler terms, the status of a moral agent also presupposes the status of a moral *patient*[108]. One useful way to capture this idea is through Robert Nozick's concepts of *moral push* and *pull:*

[104] It is clear that most AMAs certainly have financial value, in so far as they are products which are purchased or employed at some human agent's cost.
[105] Bryson, 2011: Bryson & Kime, 2011.
[106] There are some authors who would find this assumed lack of moral value and rights short-sighted (Gunkel, 2012: Coeckelbergh, 2011). However, a discussion of the possibility of robot rights, even of the relational variety, takes us too far away from our purposes here.
[107] Or at least, any moral theory which aims at the regulation of *individual* rather than say, *political* conduct. Still, even political moral theories assume that all of the agents which, say, are willing to enter into a contract, or exit the state of nature, are moral persons to which certain rights and behaviors are owed.
[108] McDermott (2008) in his reflections on what matters to a machine, captures this claim through what he calls the 'symmetry' principle of morality, which he sees as an extension of the principle of equality, and which is present in every mainstream ethical paradigm.

> Value or preciousness of persons has a dual role in my interpersonal actions. Your value generates a moral claim or constraint on my behavior toward you, because of your value, others (including me) ought to behave toward you in some ways, and not in others. Also, my value is expressed in how I am best off behaving, in the kind of behavior that should flow from a being with my value, in how that value is shown or maintained in my action. My value fixes what behavior should flow from me; your value fixes which behavior should flow towards you. Value manifests itself as a push and as a pull[109].

In Nozickian terms then, the lack of subjectivity in AMAs equates to a lack of *moral push*, or an *absence of constraints* in both a) how human agents ought to behave towards robots, and crucially b) how the moral pull of others ought to affect the behavior of the robot. Put another way, the artificial morality of an AMA—regardless of the moral theory implemented—is *necessarily* a theory of *moral pull*, since there exists no moral 'push back' flowing from the moral value of the robot itself.

This 'pushless' quality of artificial moral agency, if such a term can be used, has been addressed in the literature, and can be seen to prompt two basic conceptual moves. The first move, and one which aligns with the idea of the problem of constituency in moral theory, is to seek out a moral theory whose structure is 'pushless' in the way that seems to be required for robots[110]. This typically leads to the strong claim that this moral theory is the *only correct* moral theory for robots, even if the theory may not be 'correct' considered *sub specie aeternitatis*. The flagship argument which expounds this view is likely found in Talbot, Jenkins & Purves' article, *When Robots Should Do the Wrong Thing*, where they maintain:

> …robots cannot be agents, and their actions cannot be the subject of deontic evaluation: they cannot be right or wrong; they cannot be blameworthy or praiseworthy. Still, natural disasters can be appraised as morally good or bad, just as any other non-agentially-caused event can be appraised as good or bad…As with natural disasters, even if the deliberations and choices of robots are not open to deontological evaluation, the effects that they foreseeably bring about can be easily judged morally good or bad. If robots were to act like perfect

[109] Nozick, 1981, 401.
[110] Grau, 2006: Gips, 1995.

> maximizing consequentialists, they would bring about the best state of affairs. This is better than the alternative ways they might act, even if maximizing consequentialism is not the true moral theory...Since robots cannot be subjected to other sorts of moral appraisal—they cannot do wrong—the only moral claims that apply to robots are claims about the moral goodness or badness of the outcomes they cause. Given this, robots ought to act like perfect maximizing consequentialists (in the sense that the best state of the world is one in which robots acts like perfect maximizing consequentialists) even if consequentialism is false.[111]

Surely, Talbot and colleagues take this idea of 'pushless' morality very far, since they not only maintain that a robot is not a person, but also that it is not even an *agent*, specifically since the deterministic quality of its programming forbids the ability to act intentionally, either under the desire-belief model, or on the taking as a reason model[112]. Thus, if all robots can rightly be appraised for is the good or bad results for which they are causally responsible, then the only theory which could plausibly accommodate this type of (non) agent is consequentialism[113]. This is to say that Talbot and colleagues view the 'pushless' morality of AMAs as one which necessarily only registers impersonal moral claims regarding the goodness or badness of outcomes; any other type of moral claim is unintelligible, in virtue of the fact that robots cannot act for reasons. From there, it is but a short leap to maintain that a maximizing form of consequentialism is appropriate or optimal, since it requires only that we admit that there are certain states of the world which are morally better than others. Thus, one consequence of the pushless quality of artificial morality—and one which leans heavily on the standard view of moral agency—appears to be the vindication of consequentialism as the appropriate *foundational theory* of artificial morality. This is to say that

---

[111] Talbot, Jenkins & Purves, 2017, 260.

[112] Talbot, Jenkins & Purves, 2017, 259. "If either the *desire-belief* model or the predominant *taking as a reason* model of acting for a reason is true, then AI cannot in principle act for reasons. Each of these models ultimately requires that an agent possess an attitude of belief or desire (or some further propositional attitude) in order to act for a reason. AI possesses neither of these features of ordinary human agents. AI mimics human moral behavior but cannot take a moral consideration such as a child's suffering to be a reason for acting. AI cannot be *motivated* to act morally; it simply manifests an automated response which is entirely determined by the list of rules that it is programmed to follow." (Purves et al., 2015, 861).

[113] This is somewhat tantamount to the idea of resisting agent-centered morality in AMAs, a perspective which "…determine[s] what is right, wrong, and permissible partly at least on the basis of the individual life; his role in the world, and his relation with others. Agent-centered morality gives primacy to the question of what to do, a question asked by the individual agent, and does not assume that the only way to answer it is to say what it would be best if he did, *sub specie aeternitatis*" (Nagel, 2012, 204)

whatever the normatively relevant factors are in a given AMA's artificial morality, they must eventually be grounded in the goodness of the outcomes they produce.

Importantly, Talbot and colleagues do not initially make their argument for consequentialism at the foundational level. Rather, they defend that consequentialism ought to be the operative theory at the *normative level*, or that the normatively relevant factors which ought to matter to a machine are limited to the goodness of overall results. This leads them to make some striking revisionist claims about the ideal nature of artificial morality, mainly through the contemplation of the potential homicide of *Aunt Agatha*, a rich and frivolous elderly woman whose dwindling fortune would likely do more good in the world if her life were to end abruptly, giving what they call the *Aunt Agatha Terminator* the chance to donate a greater sum to charity. Initially, since the moral status of the killing of Aunt Agatha by a human agent likely involves certain deontological (or general) constraints involving the prohibition of causing intentional harm, but since the Aunt Agatha Terminator cannot be subject to these constraints, there is nothing morally problematic about the Aunt Agatha Terminator carrying out this dirty deed. However, since the *designers* of the Aunt Agatha Terminator *are* likely subject to general constraints of this kind, Talbot and colleagues maintain that it would be morally wrong to design a robot which acts in this way[114]. In the end, their conclusions as to the normative level of artificial morality design revolve around a theory which admits two morally relevant factors: overall goodness of results, and certain general (or deontological) constraints, the precise configuration of which depends on the degree of certainty the designer holds as to the existence or correctness of these general constraints[115].

Consequently, accepting that an AMA has no moral push does not necessarily imply accepting only 'pushless' moral theories as viable implementation strategies. Even if these theories

[114] "Even if we cannot evaluate the deontic status of the actions of the robots themselves, we can evaluate the actions of the programmers who give rise to the actions of robots…even if the world would be morally better if we designed robots to act like consequentialists, whether or not consequentialism is true, this does not necessarily make it morally permissible to create consequentialist robots"(Talbot et al., 2017, 261).

[115] Note Talbot and colleagues' choice to involve the 'full' morality of an AMA's designer as a route to curb some of the morally questionable behavior that a pure consequentialist robot exhibits. This leads them to maintain the complicated claim that the inclusion of some deontological constraints may yield the 'right' type of moral behavior, even if the underlying theory is 'wrong'. Another, perhaps simpler route, might be to maintain that embracing consequentialism at the foundational level may yield certain deontological constraints at the normative level, if the respect of these constraints indeed leads to the best results.

provide a 'best fit' solution given the ontological attributes of a machine, the further fact that a machine is a technological artefact designed by human agents may broaden the types of value that a machine ought to recognize in its moral responsiveness; and this as a pure extension of the moral responsiveness of its designer. Put another way, the arguments of Talbot and colleagues suggest that a 'pushless' morality, even if foundationally consequentialist, need not and perhaps *ought* not to act uniquely on consequentialist principles.

This passage from pure maximizing consequentialism to moderate deontology at the normative level touches upon the second type of conceptual move one can make when contemplating the theoretical ramifications caused by 'pushless' artificial moral agents. This second conceptual move is more subtle, and perhaps for this reason, more pervasive in the implicit assumptions many moral philosophers hold as to the proper design of artificial morality. In a nutshell, 'pushless' AMAs can be seen to solve the problem of the *demandingness* of a moral theory, in so far as there is nothing in these agents upon which morality could be seen to make unreasonable demands—there is no moral push, or 'value expressed in how the [agent] is best off behaving', for which morality must make room. In one rendition of this argument, the move from human moral agent to artificial moral agent likely spells the abolition of *both agent-relative permissions* and *agent-centered constraints* in moral theory, in this sense capturing the full concept of 'demandingness'. A good example of this type of claim exists in James Gips' article, *Towards the Ethical Robot*, where he contemplates the possibility of what he calls 'robots as moral saints':

> An important aspect of utilitarianism is that it is all-encompassing. To really follow utilitarianism, every moment of the day one must ask 'what should I do now to maximize the general well-being?'…Utilitarianism and other approaches to ethics have been criticized as not being psychologically realistic, as not being suitable 'for creatures like us'…Not many human beings live their lives flawlessly as moral saints. But a robot could. If we could program a robot to behave ethically, the government or a wealthy philanthropist could build thousands of them and release them in the world to help people. (Would we really like the consequences? Perhaps here again, 'the road to hell is paved with good intentions')".[116]

[116] Gips, 1995, 9.

The fact that moral theories, and especially utilitarian moral theory, prove particularly demanding for human agents is echoed in a similar paper by Grau[117], and both authors draw the conclusion that robots, precisely because the demandingness of the moral theory relates to a demandingness on the *self*, need not count as a reason against its implementation in an AMA's artificial morality[118]. In this sense, we are left with the simple idea that AMAs need not have any agent-relative permissions, and that this is likely a good thing. Interestingly, both authors seek out other reasons for which utilitarian AMAs might not be desirable, even if they appear to be ethically optimal to their human counterparts. As far as this goes, Gips seems to vaguely point to the idea that a pure utilitarian AMA may not be *acceptable*, for reasons which unfortunately he fails to expand upon. Grau, on the other hand, captures what is likely the same idea in a more sophisticated way, pointing to the types of ethical violations that a pure utilitarian AMA would likely commit in the real world. Here, he focuses on John Rawls' *separateness of persons* critique[119], and Bernard Williams' concept of 'one thought too many'[120]. Interestingly, the critiques of both authors appear to revolve around the idea that consequentialism (or utilitarianism) do not provide an adequate list of normatively relevant factors; and in this sense, even if consequentialism at a *foundational* level is both possible and ontologically correct, it is nevertheless not an acceptable or appropriate theory to implement at the *normative* level, since it seems to thwart many of the general constraints and

[117] Grau, 2006.

[118] Indeed, both authors not only allude to this concept of an artificial moral saint, but also, make allusions to the arguments of philosopher Susan Wolf. As Grau explains, "Utilitarianism…appears to require a rather different moral psychology than the sort that most people actually possess…though the life of a moral saint may be (in some ways) admirable, it need not be emulated. Such a life involves too great a sacrifice—it demands domination by morality to such a degree that it becomes hard to see the moral saint as having a life at all, let alone a *good* life" (2006, 3: Wolf, 1997).

[119] Rawls, 1971. "This [utilitarian] view of social co-operation is the consequence of extending to society the principle of choice for one man, and then, to make this extension work, conflating all persons into one through the imaginative acts of the impartial sympathetic spectator. Utilitarianism does not take seriously the distinction between persons" (24).

[120] As Grau frames it, when faced with the option of saving one's wife or a stranger from mortal peril, any justification utilitarianism can provide for saving the wife over the stranger reveals a 'deep problem' relating to the demands of strict impartiality: "…this [sort of justification] provides the agent with one thought too many; it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife" (Williams, 1981, 18).

special obligations (i.e. political rights and partiality towards one's loved ones, respectively), that many human agents both expect, and are responsive to[121].

Taking a step back then, we might deduce from the views of Grau, Gips and Talbot et al. that the principal impact a 'pushless' agent has on the fabric of morality is twofold: first, a favoritism for consequentialist moral theories at the *foundational level*, and second, a certain resistance to pure or maximizing consequentialist moral theories at the *normative level*. This latter resistance seems to take one of two forms. First, an internal form, which points to the inclusion of normatively relevant factors which do not ontologically apply to robots, but do apply to their designers. Conceiving of a robot's action as an extension of a designer's intentions, actions, or more generally, his moral agency, seems to imply the need for a type of 'mirroring effect' between the designer and the machine's moral responsiveness. This effect does not aim to ensure that *robots themselves* do no wrong, but rather, that the *designers of these robots* do no wrong, precisely by ensuring the respect of general and special obligations that may apply to human agents like themselves.

Secondly, a more external form of resistance to consequentialism seems to flow from a respect for *common sense morality*; or, that this theory by itself fails to capture the full extent of our potential considered moral judgements. Since many of the pre-theoretical intuitions which human agents hold involve more than one normative factor—which is to say, people consider more than just the overall goodness of results in their appraisal of the moral status of an act—then AMAs which do not follow suit may disrupt the fabric of moral push and pull in ways which are disagreeable or unexpected to the human agents they interact with. Put another way, in focusing only on the overall goodness of results, an AMA's responsiveness to the moral pull of its environment is likely *too narrow*. In this sense, the external quality of this latter point is derived from the idea that robots, in acting as pure consequentialists in their actions, may indeed appear to

[121] This favoritism for consequentialism is mirrored across the concrete approaches to artificial morality in the machine ethics literature (Tolmeijer et al., 2020) where most approaches, even if they accommodate certain deontological constraints, can nevertheless be seen to be structurally consequentialist. To explain this, certain authors have pointed to a correlation between an engineer's approach to rationality (the maximization of a performance measure) and the structure of consequentialism, maintaining that one reason for this theory's popularity is likely its formal familiarity to programmers (Gips, 1995 : Gabriel, 2020 : Goodall, 2014 : Dignum, 2019).

do wrong, at least from the perspective of the human agents in its external environment. At this point however, we begin to broach the thorny question of the *acceptability* of an AMA's artificial morality, and in this sense, depart from the technical limitations which concern us in this chapter. Accordingly, we will revisit this argument in chapter six, once our understanding of this concept has matured.

Still, there is a final, more *maximal* sense for which a lack of subjectivity in AMAs spells the end of 'options' in the application of moral theory to artificial morality. In this case, 'options' pertains more to the types of *agent-relative permissions* that an AMA could be seen to be privy to, or the permissibility of failing to behave in a morally optimal way if doing so would come at a significant cost to the agent. The scope of this particular view also extends past the nexus between moral theory and technical limitations, and into the purview of acceptability. However, it has a significant impact on the *place* of artificial morality in the AMA's overall agent program, and we will attempt to address this particular side of the problem here.

3.3 *The Place of Artificial Morality*

As a point of departure, it is important to note that Shelley Kagan's specific use of the term 'options' pertains, often tacitly, to any type of constraint that could permissibly prevent an agent from pursuing that action which leads to the best overall results. This particular vision of the concept betrays Kagan's utilitarian leanings, which are themselves much more explicit in some of his other work[122]. Thus for Kagan, options are really just things that get in the way of pure consequentialist considerations. Thus, interpreting the concept this way, we fall back into the first sense of 'options', where 'pushless' artificial moral agents are at least *foundationally* consequentialist since this is the 'best fit' theory, given their particular ontological characteristics. Thus, if we are to understand this more maximal sense of the implications of 'pushless' morality, we must divorce ourselves from the perspective of Philosophical Utilitarianism[123]. We can achieve

[122] Kagan, 1989: 1994.

[123] "…what I will call 'philosophical utilitarianism' is a particular philosophical thesis about the subject matter of morality, namely the thesis that the only fundamental moral facts are facts about individual well-being. I believe that this thesis has a great deal of plausibility for many people, and that, while some people

this by asking, in the case of 'options' understood as agent-relative permissions, what could be considered as a 'cost' to an agent that does not have any moral push, self-interest or value?

In a certain sense, we might be able to identify one such 'cost' if we look carefully at a portion of Gips' argument cited above: "If we could program a robot to behave ethically, the government or a wealthy philanthropist could build thousands of them and release them in the world to help people"[124]. Resurrecting a term from previous chapters, it would seem that for Gips, the *purpose-oriented ontology* of a 'moral saint' AMA is (or must be) 'to help people', in so far as the AMAs are built to achieve this purpose in one, or likely many *Umwelts*. These ethically optimal AMAs, in other words, are *not* designed to drive people from A to B, to assist in medical operations, to decide the chances of recidivism for a given criminal, or to fight in armed conflicts. Instead, the extensional restriction of Gips' Moral Saint AMA appears to be *subsumed by its morality*, generating what we might call a Philanthropotron3000™, whose only purpose is to do good in the world, or bring about the best consequences. Quite clearly, this directly conflicts with the tool AI assumption which underpins every current artificial moral agent, or every level 4 explicit ethical agent, where morality is seen as a 'patch' on the otherwise practical pursuit of a given human agent's end. This, rather obviously, has quite serious implications for the *place* of artificial morality, and we can find two such examples of how this occurs in Derek Leben's work *Ethics for Robots: How to Design a Moral Algorithm*[125]. The first case occurs in Leben's exploration of the use of the harm principle in artificial morality design:

> You go to the local gas station, and ask the robot clerk for a pack of cigarettes. The robot refuses. You ask again, but the robot politely explains: 'I'm sorry, but selling you cigarettes reduces the likelihood of your survival by 0.00004 percent. Therefore, this action is restricted by my ethics engine[126].

Another such case occurs in the exploration of his favorite criterion of rightness, the maximin principle:

---

are utilitarians for other reasons, it is the attractiveness of philosophical utilitarianism which accounts for the widespread influence of utilitarian principles" (Scanlon, 1982, 108).

[124] Gips, 1995, 9.

[125] Leben, 2018.

[126] Ibid., p. 84.

> Every Saturday afternoon, your robot personal assistant buys you groceries and brings them back to your apartment. On one of these Saturday afternoons, you notice that the robot only brought back a third of the groceries on your list. You say, 'Robot! What's the problem? Why didn't you buy all the groceries?'. The Robot sheepishly responds, 'I did, but then I passed two homeless men on the way here'[127].

Leben goes on to maintain that these types of agents, which he calls *morally superior robot villains*, constitute the correct way to design artificial morality, even if this thwarts the preferences or moral expectations of the individuals these agents interact with, or smacks of a particularly pungent form of technological paternalism. We will address these sorts of claims in chapters V and VI. Here, our aim is rather to point out the interplay between agent-relative permissions and the place of artificial morality. In this sense, it would appear that the place of morality that Leben imagines, in both of his illustrations, is all-encompassing: the artificial morality of his AMAs is never *dormant* and is constantly influencing their action selection. Correspondingly, Leben must maintain that every potential context within the *Umwelts* of these AMAs (a gas station, and a neighborhood, respectively) constitutes a *context of ethical salience*; or, that every action these AMAs perform requires ethical responsiveness.

How does Leben arrive here? It appears that Leben makes a sort of argument from generalization, assuming a lack of agent-relative permissions in an AMA's artificial morality, and using Peter Singer's famous drowning child thought experiment[128] as fodder: "If you agree that it would be morally wrong to allow a child to drown in a shallow pond to save your expensive clothes, then it appears inconsistent to allow children to die of malaria when you could be donating money to charities…And a robot programmed to save children in dying ponds would immediately start giving your money away as soon as it has any control over your bank account"[129]. One way of interpreting this is the following: given that Leben's Singerian robot does not have any agent-relative permissions allowing it to forgo saving the child in order to avoid some cost to it (ruining

[127] Ibid., p. 42.
[128] "If I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing" (Singer, 1972, 837).
[129] Leben, 2018, 42.

its expensive clothes seems unrealistic), this same robot does not have any agent-relative permissions *anywhere*, including its financial decisions as to its user's bank account, since the place of its utilitarian artificial morality is omnipresent.

In response, we might take issue with three points. First and trivially, it is perhaps not realistic that any current AMA would be ontologically equipped (both in terms of hardware and software) to save a drowning child *and* to have access to an individual's bank account, since this presupposes a very *universal* type of AMA indeed. Secondly, Leben's robot seems to require a type of moral *proactivity*, wherein ethical constraints guide not only action selection, but *goal generation*. This type of artificial morality is rather atypical in the machine ethics literature, since most approaches aim only at constraining practical action in ways which align with what morality requires[130]. Lastly, Leben appears to presuppose a *uniformity of moral pull* across both decision contexts, aided by the power of the logical consistency of utilitarian principles, and a lack of agent-relative permissions. In other words, *every* situation which this robot encounters will contain as much ethical pull as that which the drowning child exerts on the 'smartly dressed business robot', precisely because there is no *moral push* that regulates the cost to the robot, or which separates every-day decision-making from moral *dilemmas* or moral *emergencies;* or put another way, no moral push which separates right or permissible actions from *supererogatory actions*. In this sense, the robot is always morally required to strictly or *maximally*[131] apply whatever moral principles flow from a theory's normatively relevant factors (maximize welfare, do no harm), even if this gets in the way of its practical purpose, e.g. to sell cigarettes, or to buy groceries for a particular user.

Straightforwardly then, the cost of such a maximal view of 'pushless' morality is not the robot's own self-interest or commitments, but rather, the *practical purpose* the robot is meant to

[130] Dennis & Fisher, 2018.

[131] This use of 'maximizing' can be contrasted with a 'satisficing' account of individual morality (or rationality), which underpins the idea that "…acts that are less than the best (most beneficent) possible as sometimes good enough and so not morally wrong even apart from any sacrifices a better (more beneficent) act might require from the agent" (Slote, 1985, 142). Indeed it would appear that Leben, owing perhaps to his Rawlsian leanings, presupposes a maximizing account of individual rationality which mandates that the agent select the optimal or 'best' option according to the relevant criterion of rightness, rather than actions which are 'good enough'. Importantly, both the concept of supererogation and satisficing morality are often seen to figure in common sense morality (Slote, 1985).

perform in an *Umwelt*. Philanthropotron3000™ while something of an artificial moral saint, cannot achieve any practical goal with consistency in its environment. This bold claim likely needs some refinement. Indeed, while Leben's argument leaves much to the imagination, on at least one reading of his view, he seems to be making a very decisive point, since if strict adherence to moral theory has a continuous, lexical priority over the achievement of practical ends, two undesirable scenarios seem possible. First, as Leben demonstrates, there is the possibility that the demands of morality will frustrate the practical purpose for which the AMA is designed, yielding a robot that does not do its 'job' effectively. This can happen either a) because the robot is *too* responsive to the moral pull of its *Umwelt*—such as when a robot considers that an action which results in a 0.00004 percent drop in a human's survival rate is constitutive of *harm* to that human, and thus forbidden according to its artificial morality, or b) because the robot is responding to the wrong[132] sorts of normatively relevant factors, or not *enough* normatively relevant factors—such as when a robot personal shopper considers that the only factor of normative importance is the maximization of the minimum share of society, and thus redistributes its user's groceries to the less fortunate.

Secondly, the continuous, lexical priority of morality over practical ends seems almost to *require* a universal agent, of the type imagined by Gips. We have seen the rather *proactive* side of this claim with Leben's 'smartly dressed business robot', who moves effortlessly from saving the child to managing its user's finances, thus seeking out those *Umwelts* within which it is able to do the most good, or bring about the best results. To this end, the *Aunt Agatha Terminator* of Talbot and colleagues appears to be similarly universal, since it is not clear for what particular purpose such a robot would be constructed. It is also important, however, to imagine the *prohibitive* side of this claim, pointing to the idea that there are certain *Umwelts* within which 'moral' robots refuse to engage. Leben captures this point nicely with his subsequent discussion of robot soldiers, who, operating on the maximin principle, refuse to engage in a war of aggression[133].

---

[132] We intend 'wrong' here as internally or externally unjustifiable—rather than wrong *sub specie aeternitatis*—in line with our analysis of the previous section.

[133] "Soldier-bots must also be capable of recognizing which entities constitute an 'active threat', and have access to a large database of how effective various attacks have been in preventing future threats. A contractarian ethics engine will respect the primary goods of enemy combatants, since every agent is equally likely to be one's political enemy from the original position. Thus, if enemy combatants pose an active threat, the ethics engine will recommend the minimal amount of force necessary to respect the health and survival of all relevant parties. This also necessarily prevents soldier-bots from being used in wars of aggression" (Leben, 2018, 145).

If we assume, as these authors likely have, that these AMAs are embodied and somewhat humanoid, it may seem laudable that they would stand their ground, or 'walk away' in the face of morally dubious action options. However, it would seem that this romantic form of anthropomorphism tricks us into creating 'practical duds', or machines which, under certain circumstances, cease to be *agents simpliciter*. To illustrate, it seems fair to assume that Leben's pacifist robot soldier would simply shut down when faced with engagement in a war of aggression, since the material conditions of its existence likely prevent it from say, running away and joining an Ashram. Similarly, a cashier robot endowed with a strict form of the harm principle would likely find that *no* potential purchase constituted an ethically viable option for a customer—since junk food, lighters, gas and the plastic bags they come in, all pose a potentially harmful threat of some kind—and thus refuse to sell anything. While these constitute rather comical examples, they nevertheless frame the idea that if given absolute lexical priority, a robot's artificial morality may not only frustrate the robot's purpose, but perhaps confound it entirely. To this end, if the designer's pursuit of morally dubious ends proves troubling—as is likely the case at least with autonomous weapons and wars of aggression—it is surely not the place of *artificial morality* to correct the moral quality of these pursuits to the point of stopping them entirely[134].

3.3.1 *Three Visions of the Place of Artificial Morality*

In general, and in light of our analysis thus far, we may conclude that a robot may not have any *moral push* flowing from its value as a person, but it might need some degree of *practical*

[134] To this end, it would seem that Leben relies on an assumption common among some machine ethicists of the philosophical persuasion, namely that the designer has but *one* opportunity to decide the content and structure of his robot's artificial morality. A similar assumption is made in (Barghava & Kim, 2017) and Talbot and colleagues' argument, where they maintain that "Creating a robot is a single act, and if creating that robot violates one of the creator's duties…that duty is the product of the conjunction of duties we expect the robot to violate"(2017, 268). This claim is surely misguided. While it is commonplace to maintain that robots will encounter unforeseeable or unpredictable decision-contexts—thus posing a problem for an 'autonomy as provision' approach to the design of artificial morality—it does not follow that the robot is then impervious to any further tinkering or 'updating', acting suddenly as an impenetrable autonomous agent whose practical principles cannot be modified. It certainly seems empirically possible to update the artificial morality of an AMA at least periodically, and if this is plausible, this casts serious doubt on whether machine ethicists must 'get it universally right the first time', and thus the precautionary force of their espousal of a particular moral theory.

push if it is to avoid confounding its very purpose. Or, the place of artificial morality cannot be lexical and continuous if the resulting artificial agent is to efficiently and reliably satisfy its extensional restrictions, or purpose. One way to achieve this is to incorporate a form of practical push into the artificial morality of an AMA, an option we will explore in chapter VI. Another, more basic solution to this problem however, is to restrict the place of artificial morality to particular types of decision contexts. In a related sense, Jason Borenstein and Ronald Arkin discuss the (moral) permissibility of 'nudging humans to become more ethical', a situation which might occur when say, a companion robot notifies a parent that her child has been sitting alone watching TV for a long time, thus implying that family time is needed. Importantly, the concept of a 'nudge' only entails the *suggestion* of ideal behavior to human agents, rather than the ostensible enforcement of it, as Leben seems to do. Still, even in this weaker form, Borenstein and Arkin remain trepidatious: "…whether a robot should be permitted to perform a particular 'nudging' act will be contingent in part on its level of familiarity with a user. This implies that a robot would need to be sophisticated enough from a technical perspective to distinguish between different human beings and possess enough situational awareness to discern when performing certain types of behaviors is appropriate…[however] it does not necessarily follow that what we would permit in human-human interaction should be allowed within the context of [Human-Robot Interaction]"[135]. Arkin and Borenstein hold not only that this level of computational sophistication is possible[136], but also that it is necessary to create certain barriers to intrusion on a rational agent's choices[137]. Here, we can see the idea that the 'practical push' of machines such as companion robots likely involves a certain degree of allegiance to its user[138], further bolstering the link between a lack of agent-relative permissions and a lack of moral responsiveness. With this distinction between 'appropriate' contexts for moral responsiveness then, we appear to circle back to Virginia Dignum's final criterium of a top-down approach: the need for "…deliberation capabilities to decide whether the situation is indeed an ethical one"[139].

[135] Borenstein & Arkin, 2016, 36.
[136] Arkin et al., 2003.
[137] This becomes clear in their discussion of different types of nudging frameworks, and their correlation with moral paternalism and bioenhancement (Borenstein & Arkin, 2015, 38-41).
[138] Lin, 2016.
[139] Dignum, 2019.

Indeed it is likely that what Dignum intends as an 'ethical situation', though slightly different from Borenstein and Arkin, reduces to areas of significant ethical pull[140], such as lethal accident scenarios in autonomous vehicles, lethal combat in military robots, or significant value trade-offs (i.e. beneficence v. Patient autonomy) in the case of robotic healthcare assistants. This tracks the more machine-ethics-centric distinction between mundane and complex cases, where only the latter are seen to require an 'optionless' account of artificial morality[141]. The main reason for this distinction relates to the concept of *normative convergence*, as Tolmeijer and colleagues argue in their extensive survey article on approaches to artificial morality:

> …in ordinary life situations in which one is confronted with extremely difficult ethical decisions or runaway trolleys are exceedingly rare. In many domains, moral dilemmas are unlikely to arise or be of much import, and there is widespread convergence (not only among the folk, but experts, too) on what constitutes adequate moral behavior…[142].

The concept of normative convergence then leaves us with a two-tiered vision of ethically salient contexts. First, there appear to be contexts of ethical salience where the robot must exhibit some form of moral responsiveness, but critically, the form this moral responsiveness ought to take is widely agreed upon. For instance, it is plausible that there exists significant convergence concerning the general idea that a robot should not deliberately harm a human being, unless this harm is somehow unavoidable; and more contextually, that a robot doctor should not deliberately harm a human patient, in virtue of this general sentiment but also perhaps in virtue of its adherence to the Hippocratic oath. Secondly, there appear to be contexts of ethical salience which pertain to complex cases where rational and suitably informed human agents, or even experts, can still be seen to disagree about what morality requires of an AMA's responsiveness[143]. Cases such as an autonomous vehicle's 'deciding how to crash', or a robot doctor's 'deciding who receives a ventilator' likely constitute such instances. Put bluntly, there are contexts where moral behavioral

[140] What we intend by areas of 'significant moral pull' are simply decision contexts wherein human agents hold strong (or perhaps equally strong) moral claims over the behavior of the AMA, in virtue of the significant effects an AMA's decision will have over their welfare, or some other morally relevant feature.
[141] Tolmeijer et al., 2020: Himmelreich, 2018: van Wynsberghe & Robbins, 2019: Nallur, 2020.
[142] Tolmeijer et al., 2020, 10.
[143] Gabriel, 2020 : Tolmeijer et al., 2020.

standards are generally accepted and clear, and others where opinions diverge. Given this distinction, what is the proper place of artificial morality?

In terms of pure published material in machine ethics, there seems to be something of a tacit attraction to complex or dilemma cases in both the theoretical exploration and computational formulation of artificial morality[144]. However, it is unclear whether this fixation on complex cases is intentional and conscious, or whether it is just the by-product of an attraction for the exciting theoretical opportunities these sorts of cases present. This ambiguity is further frustrated by the various types of paradigms which provide normative frameworks for the design of artificial morality. There are those machine ethicists that see the role of artificial morality as one of balancing competing *values* inherent to an AMA's *Umwelt*[145], such as 'autonomy', 'efficiency' and 'safety'; and as we have seen, those that rather see it as an application of computational social choice theory, or various moral theories from the philosophical tradition. Since each of these approaches track fundamentally different features at the normative level—user preferences, societal preferences, the morally relevant features of a given moral theory, etc.—and influence the structure of artificial morality in different ways—from 'counts as' norm violations in value programming[146], to aggregation[147], to moral deliberation—each particular configuration yields its own account of the proper place of artificial morality. Add to this the incredibly divergent claims concerning the variety of *Umwelts* which could plausibly require a form of moral responsiveness from AMAs[148], regardless of the type of theoretical and computational decision procedure implemented, and we appear to be left with a mosaic of places in which artificial moral agency ought to inform AMA behavior, and an equally irreducible account of what moral responsiveness ostensibly entails. Accordingly, we will limit ourselves to the application of moral theory in top-

[144] Lin, 2016 : Bonnefon et al., 2016 : Conitzer et al., 2017 : Pereira & Saptawijaya, 2007 : Evans et al., 2020 : Keeling et al., 2019 : Santoni de Sio, 2017 : Goodall, 2013 : 2014 : 2019.
[145] Dignum, 2019, 39-41: Miceli & Castelfranchi, 1989: Gerdes & Thornton, 2015: Evans et al., 2020: Himmelreich, 2018: Kearns & Roth, 2019.
[146] Aldewereld et al., 2010: McLaren, 2006: Malle & Scheutz, 2018.
[147] Conitzer et al., 2017 : Noothigatu et al., 2018.
[148] From those that maintain that 'everyday decisions' require moral responsiveness (Scheutz, 2016) to those that deem it necessary only in situations where lethal consequences are at play (Lin, 2016: Keeling, 2020).

down accounts of artificial morality, and attempt to delineate what we will call the *narrow*, *moderate* and *wide* view on this issue[149].

Given the analysis of this section, we will not need to look far to discover the *wide* view of the place of artificial morality. Indeed, Derek Leben has given us a very clear vision of what this view entails. In this sense, a subscriber to the wide view holds that the proper place of artificial morality is omnipresent, which is to say, the wide view supports the idea that artificial morality *is* the agent program which moves an AMA from perception to action in its environment. Again, holding this view, we must accept that every decision context in an AMA's *Umwelt* counts as a context of ethical salience, and thus that moral responsiveness is continuously required. However, in accepting the wide view, we are not necessarily tethered to the maximalist and 'optionless' view held by Leben. In other words, even if the place of artificial morality is omnipresent, this does not logically entail either a) that a particular moral theory constitutes the 'right' moral theory to implement, even if it disrupts the moral fabric of common sense morality, or b) that the theory implemented must leave no room for agent-centered constraints, or agent-relative options, particularly of the *practical push* variety. Put bluntly, the wide view does not automatically lead to the design of a Philanthropotron3000™. To this end, Ronald Arkin seems to expound a wide view of the place of artificial morality in his work on autonomous weapons, since the Law of Armed Conflict, the Rules of Engagement, and mission-specific guidelines are all implemented into a top-down decision-procedure which regulates *all* of the agent's behavior[150]. In this sense, Arkin's adoption of the wide view should not be that surprising, since it seems likely that theatres of war and instances of armed conflict (the *Umwelt* of these types of agents) constitute ethically salient contexts[151].

[149] As a point of clarity, all three visions which we will explore in this section pertain to what Bauer (2020) has called 'small scale interactions' in the study of machine ethics, which pertains to "…relatively small groups of interacting agents (at least two), which might involve human agents, [where] we want to specify which norms artificial moral agents follow in their interactions" (2020, 264). These differ from the 'large-scale' ethical governance of sociotechnical systems, such as the types of artificial agents which operate on the stock exchange (Crawford & Calo, 2016: Chopra & Singh, 2018).

[150] Arkin, 2009.

[151] In this respect, the wide view understood from the angle of an agent program can coincide with a narrow view understood from the angle of morally relevant features, since criterions such as 'life and death decisions [being] made on a daily (or hourly) basis"(van Wynsberghe & Robbins, 2019, 723) can occur frequently in some *Umwelts* (i.e. warfare or medicine), therefore requiring constant moral responsiveness, while being comparatively rare in others (i.e. autonomous vehicles), requiring a more narrow

On the other hand, a *narrow* view of the place of artificial morality, as the term implies, is highly restrictive in its use of artificial morality, and necessarily presupposes the existence of two types of agent program in an AMA's architecture. This view coincides with an account of artificial morality which aims at the resolution of complex, or *dilemma cases*, and thus requires the deliberation capacities needed to detect an ethical situation, in line with Dignum's criteria[152]. At a theoretical level then, a thorough definition of what can count as a moral dilemma must also be provided. A good example of one such definition is provided by Geoff Keeling in his work on autonomous vehicles, where he defines the place of artificial morality as a response to the *moral design problem:*

> Suppose a driverless car encounters a situation where (i) inflicting death or harm on at least one person is unavoidable; and (ii) a choice between how to allocate death or harm between different persons is required. What does morality require of driverless car manufacturers in these cases? How, morally, should these cars be programmed to allocate death or serious harm between different persons?[153]

For Keeling, it appears that the place of morality not only coincides with the resolution of trolley-type dilemmas—a rather general conviction within the field of autonomous vehicle ethics[154]—but further, that it involves selecting a moral theory (or normative approach) which a) tracks the morally relevant feature of 'harm', and b) provides an action-guiding recommendation for *harm allocation* in unavoidable collision scenarios. Further, what Keeling calls the moral design problem likely coincides with what Tolmeijer and colleagues call a 'complex case', wherein rational and informed human agents can still disagree about what morality requires. Notice then that the place of artificial morality does not include a number of potential ethically salient contexts which an autonomous vehicle might encounter: deliberation as to the potential violation of traffic

---

responsiveness. Still, the type of wide view with which we have taken issue in this chapter is typified by a correspondingly wide account of what can count as an act of harming, one which extends well past lethal harm.

[152] We will not discuss what types of factors indicate the presence of an ethically salient context here, since this discussion is addressed in chapter VII.

[153] Keeling, 2018, 1.

[154] Lin, 2016: Gerdes & Thornton, 2015: Goodall, 2014: Keeling et al., 2019: Evans et al., 2020: Santoni de Sio, 2017.

regulations (such as speed limits), deliberation concerning the appropriate amount of risk to which the vehicle is able to submit its passengers, or deliberation concerning optimal route planning via socio-ethical criteria such as low-income neighborhoods, 'eco-friendly' routes, or comfort to the passenger, among others. All of these things, if they are to matter morally to the vehicle at all, are then punted to the tactical planning mechanisms of the vehicle, or its principal and practical agent program. In this sense, while the vehicle's ethical responsiveness in these situations may be very *broad*—including things other than the maximization of the goodness of results—the place of this behavior is nevertheless very *narrow*, in so far as it is relatively unlikely that an autonomous vehicle will be able to exhibit moral responsiveness very often in the practical perusal of its environment.

Finally, we are left with what we have called the *moderate* view of the place of artificial morality. This view, unsurprisingly perhaps, is somewhat elastic compared to its more extreme counterparts, depending heavily on the type of *Umwelt*, and the type of AMA in question. At a very abstract level however, the moderate view often challenges the supposed *normative convergence* around the moral behavioral standards endemic to the performance of a particular task or role; this underpins the idea that artificial morality ought to regulate an agent's behavior whenever there can be observed conflicts between the expectations—or the moral claims—of the human agents in the agent's environment, or in a correspondingly large sense, in terms of societal expectations, or political questions of justice and fairness. One such example, keeping with the theme of autonomous vehicles, can be seen in Johannes Himmelreich's defense of the ethics of mundane cases in autonomous vehicle decision-making[155]. Himmelreich advances two ethico-technical arguments for the ethical importance of contexts such as approaching a crosswalk with limited visibility, making left turns, or navigating through busy intersections: what he calls the problem of *specificity* and the problem of *scale.* Specificity relates to the *intuitive* quality of the human decisions which inform risk mitigation contexts, such as those which arise when a human driver considers whether to stop for a pedestrian that appears to want to cross the road. Scale relates the idea that the uniformity of programming across autonomous vehicles might turn little assumptions in mundane scenarios into general policies, the ethical ramifications of which will be

[155] Himmelreich, 2018.

felt more often than those of unavoidable crash scenarios. Importantly, Himmelreich uses these critiques to point to multiple sets of trade-offs in moral values (such as economic freedom, intellectual property rights, mobility and environmental impact) which are ostensibly *missed* by narrow views such as Keeling's moral design problem[156].

Under a strict reading, it would seem that Himmelreich is questioning the focus on (or perhaps even the need for) artificial morality's applications to unavoidable collision scenarios[157]. Here then, he would be advocating for a moderate view of artificial morality which likely does not include the use of standard moral theories in unavoidable collisions. Value trade-offs and more politically oriented moral frameworks ought to inform the vehicle's decision-making, rather than the standard 'trolley-type' moral theories which are meant to underpin individual moral decisions[158]. This is one way to approach the concept of a moderate view, one which from a critical perspective, aims to show that narrow views are not sufficient, and in so doing, typically proposes some revision on the types of theories, or in some cases the kinds of normative framework, which underpin the narrow view.

Another account of the moderate view however, entails an escape from the binary assumptions of the narrow view. To illustrate, it may be possible that in the case of autonomous vehicles, both unavoidable collisions *and* mundane contexts carry ethical salience, and critically,

---

[156] In detail, he identifies three distinct ethical concerns: "First, the optimization problem to make autonomous vehicles as safe as possible puts at stake issues of economic freedom and intellectual property rights. This is an internal value conflict that arises in the process of achieving global safety optima. Second, further values – such as mobility, environmental impact, or values in urban design and traffic planning – might conflict with safety. How these concurring values are balanced against each other is an important ethical question. Third, existing legal frameworks give rise to perverse incentives. Adjusting the framework and mitigating against these incentives is a delicate issue because the effects of legal changes are potentially widespread." (Himmelreich, 2018, 16).

[157] To this end, Himmelreich also seems to toy with both sides of Bauer's (2020) distinction between small-scale and large-scale questions in machine ethics, pointing to the need for increased attention towards the latter.

[158] "A political approach takes as its starting point the diversity of views and values that are the predicament of any political community…This political approach contrasts with the approach of moral philosophy that is taken by trolley cases. Trolley cases make no room for such pluralism by aiming to elicit an individual's decision. A trolley case prompts us to make an individual choice when what we in fact face is a social choice. What seems needed is a kind of compromise to overcome disagreements over issues of value. Insofar as we value the moral diversity of our political community, it should be recognized that autonomous vehicles pose primarily a political problem, not a moral one". (Himmelreich, 2018, 9).

that each type of ethically salient context requires its *own* moral theory or normative structure, generating *two separate decision procedures* which trade in the dominance of the vehicle's tactical planning, depending on the presence of certain action options or contextual features. Indeed, viewing things this way, it would seem that Leben's 'smartly dressed business robot' would benefit from such a configuration, since this would avoid the *overgeneralization* of a utilitarian criterion of rightness from which this robot suffers, and thus its distinctly maximalist approach to artificial morality. Put another way, a moderate view may be able to accommodate differences in the *strength of moral pull* across different decision contexts, in ways that both maximalism and minimalism cannot. In this sense, rather than assuming, under a wide view, that morally superior robot villains are a necessary means—or a negative external effect—of ensuring that the right moral theory is applied in dilemmatic or morally dangerous places of an AMA's behavior, it is certainly conceptually and technically possible to build an AMA with *two* criterions of rightness, and thus two fundamentally different forms of contextually-sensitive moral responsiveness[159]. This is but one of the relatively exotic benefits of the application of moral theory to non-human agents: a split (or even multiple) 'pushless' moral personality which addresses contextual changes in what morality requires.

In sum, these three visions of the place of artificial morality serve to challenge the distinctly maximalist and wide view of artificial morality, one which consciously or not, assumes that the proper place of moral theory in AMA design revolves around the creation of universal moral saints which are unable (or computationally unwilling) to serve many practical purposes. In effect, each alternative vision provides its own account of the proper division between moral pull and practical push, while all accounts provide a feasible framework for level 4 explicit ethical agents.

[159] We might object that a) this sectioning-off of the contexts of an AMA's *Umwelt* may seem to ready us for a casuistic infinite regress, and b) that these multiple personalities resemble deliberation *across* multiple criterions of rightness, which we held in previous sections to be relatively computationally intractable. To the latter objection, we might say that there is a salient conceptual difference between deliberation across principles in a *single* decision context, and deliberation using different sets of principles *depending* on the context. Without making too many generalizations, it would seem that the computational load required to detect whether a vehicle is about to cross the path of a pedestrian or *collide* with that pedestrian, is lesser than the load it would take to evaluate the interests and justifiable reasons for action for *every* human agent in the traffic environment, including that pedestrian. To the former objection, we might say that this infinite regress is avoidable if a given decision procedure can accommodate a measurement of the strength of an individual's moral pull (or moral claim). We will propose such a decision procedure in chapters VI and VII.

4. *Conclusion*

In this chapter, we have addressed the engineering and philosophical conceptions of the three main approaches to artificial morality: top-down, bottom-up and hybrid. Most generally, these approaches can be distinguished by their relative dependence on the use of (moral) *theory*: top-down approaches typically constitute computational models of an *interpretation* of a moral theory, while bottom-up approaches find ideal behavior through *a*theoretical means, via reinforcement or mimicry. More specifically, these approaches can be distinguished in terms of the *source* of moral content, where top-down approaches rely on an explicit moral theory, and bottom-up approaches rely, to various degrees, on the (moral) wisdom of the crowd. These positions constitute the basis of what we have called the maximalist and the minimalist view of machine ethics, respectively. The maximalist's use of moral theory both as an agent program and as a source of moral content is burdened by two problems: the *constituency* of a moral theory, which typically demands ontological features and moral dispositions, attitudes or mental capacities which current artificial moral agents do not possess, and the *place* of moral responsiveness, which may restrict the choice of moral theories which are appropriate to implement, or the overall role of moral decision-making in the larger practical agency of the AMA. The minimalist, however, in his mobilization of the wisdom of the crowd, runs into problems when a) the nature of the content collected (beliefs, intuitions, preferences) is uncertain, or when b) the decision principles generated appear to indicate bias, incoherence, irrationality, or otherwise ethically dubious action-guiding recommendations.

Focusing our attention on top-down approaches, we questioned the viability of the *simple thesis*: that a computational decision procedure can be structurally and functionally independent from the type of values, principles, or moral theories we seek to implement. Exploring the basic structure of a moral theory, we deduced that a strict reading of the simple thesis is false; while we can interpret moral theories to better fit the confines of computation, there exists certain factors—especially in terms of the theoretical decision procedure—which ought to be emulated at the computational level, at least if a 'tight' interpretation is desired. From there, we explored one particular technical limitation that plagues this nexus between moral theory and computational implementation: the lack of subjectivity in artificial moral agents. This led us to conceive of any

type of artificial morality as a theory of moral pull, and consequently, to conceive of AMAs as 'pushless' moral agents. At a philosophical level, the ramifications of pushless morality seemed to manifest themselves in an attraction for consequentialist theory at a foundational level of artificial morality, but nevertheless engendered a certain resistance for this type of theory at the normative level. Pushless morality also lead us to explore the idea of robotic moral saints, or 'optionless' moral agents whose artificial morality is both lexically prioritized and continuous across their action. We saw that the creation of these agents, while initially quite desirable, could easily lead to a maximalist view of artificial morality which negated the importance of the *practical push* of an AMA's purpose-oriented ontology, and a great number of our considered moral judgements about ideal AMA behavior. Finally, we enumerated three visions of the place of artificial morality, which accorded relatively more or less of a place for the practical push of an AMA.

CHAPTER

5

Acceptability & Artificial Morality

In the first section of this thesis, a great amount of attention was spent in understanding the ontological ramifications of artificial agents understood as technical artefacts; the most basic interpretation of which leads us naturally to two conclusions: first, artificial moral agents did not arrive *ex nihilo*, and second, that they are not organic or natural entities, issuing for instance, from a process of biological evolution. In the last chapter, we added a further ramification of AMAs as technological artefacts, their lack of *moral push*, this time flowing from the fact that AMAs cannot (currently) meet the conditions of the standard view of moral agency. In a sense then, the fact that artificial moral agents are created, designed and used by humans, and further, that they do not have any moral value flowing from their status as explicit ethical agents, seems to point to a general vision of AMAs as artefacts who are built to *serve*; serving either the interests of individual humans or societies, or perhaps, the interests of non-human entities that humans generally hold to be valuable, such as sentient animals, or the environment.

While this decidedly humanist vision of artificial moral agents has come under philosophical scrutiny in recent years[1], it remains clear that the intentions behind the design of most current technological artefacts remain human-centric, whether acting specifically as exotic tools for the pursuit of human ends, or enhancing cooperation and efficiency in larger socio-technical systems. Importantly, this humanistic perspective of artificial moral agents has a significant impact on the types of expectations and evaluative attitudes we hold towards the behavior of AMAs, and in a certain sense, their overall impact on the human social sphere. Most principally, it would seem that the humanist intentions behind technological development provide grounds for the importance of the *acceptability* of these technologies; and thus, that the expectations, preferences and attitudes individuals hold vis-à-vis a certain technology ought to be at least considered in decision-making at the design level. Put in more practical terms, if technologies are meant to be used by individuals, then it appears important to ensure that they are indeed useful to these individuals, by investigating the types of intentions individuals have towards their use, or the types of preferences they hold in regards to the scope and details of their functionality. This coincides with what is likely the most basic concept of acceptability, one which aims to ensure that "…agents [are] designed to fit well with how people actually work together…assur[ing] that effective and natural coordination, appropriate levels and modalities of feedback, and adequate predictability and responsiveness to human control are maintained"[2].

Nevertheless, given the relative novelty of many forms of AMAs, and in many cases, the unprecedented functions that they can perform, deciphering the substantive content of what we expect of their personality and impact can be challenging in the extreme: a task which at times seems equivalent to guessing what people might expect from a benevolent alien species. Perhaps in recognition of these difficulties, many of the more philosophically inclined machine ethicists have turned to various myths, fables and stories to elicit a type of normative narrative which could provide insight in these pursuits; where fictional characters such as Frankenstein's monster, Pygmalion, Prometheus, H.A.L. 9000[3], and Asimov's many robots[4] stand as particularly cogent candidates. As Alexei Grinbaum, a particularly fervent supporter of this approach claims:

[1] Coeckelbergh, 2020.
[2] Bradshaw et al., 2004, 365.
[3] Dennett, 1998a.
[4] Gips, 1995.

> Considering technology within the prolongation of ethical traditions entails giving myths a fundamental importance. I do not see any solution other than to afford algorithms a place among the narratives that have contributed to the formation of the bedrock of our civilization and culture. Since, if these technologies are new, the moral questions they pose often echo the interrogations of other contexts and eras[5]…

In this sense, if technologies such as autonomous vehicles constitute such an exotic object of assessment, to the point of leaving us speechless or in want of what to expect, the same narratives that afforded us a moral education in childhood—or a compelling glimpse at a possible world—may guide us through the task of deciphering the uniqueness[6] of technological artefacts. However, even if stories such as Mary Shelley's *Frankenstein* afford us a potentially useful *general* point of entry into the evaluation of non-human agents capable of acting for good or for evil, the type of conclusions we can draw from this method remain quite divorced from the day-to-day impact that technologies such as autonomous vehicles or robotic healthcare assistants will have for human agents. In one sense, this is because many of these narratives derive their principal allegoric potential from the relationship between a machine and its *designer*; and in another sense, fail to account for the role that a purpose-oriented ontology plays in the expectations human agents hold for a specific type of AMA. Accordingly, while it may be that human agents, whether for instance addressing the moral risks of H.A.L. 9000 or virtual assistants such as Siri, hold a concern for the degree of beneficence or control these non-human actors exhibit across their behavior and decisions, it does not follow that what David Bowman and Frank Poole expected from H.A.L. meaningfully corresponds to what a standard human user expects from her smartphone, unless she too happens to be planning an expedition to Jupiter. Consequently, if we are to fully embrace the humanistic intentions of technology, and thus take acceptability seriously, philosophical inquiry

[5] Grinbaum, 2019, 7. Translated by the author, the original text reads: "Penser le numérique dans le prolongement des traditions éthiques signifie donner une importance fondamentale aux mythes. Je ne vois pas d'autre solution que de ménager aux algorithmes une place au sein des récits qui ont contribué à former le socle de notre culture et de notre civilisation. Car, si les technologies sont nouvelles, les questions morales qu'elles posent font souvent écho aux interrogations surgies dans d'autres contextes et à d'autres époques."
[6] It would seem nevertheless that Grinbaum assumes a negative answer to the Uniqueness Debate exposed in the first chapter, one which might minimally maintain that there is a significant correlation between traditional ethical problems, and those ethical problems posed by emergent technology.

alone may not provide a sufficiently robust account of what matters (morally) to human users. Instead, empirical and investigative methods must fill in the theoretical gaps left unanswered.

However, even if it can be granted that the consideration of human approval and the mapping of human expectations is a necessary step in the design process of an AMA, we are again left with the harrowing task of deciphering exactly which questions to ask, and exactly which expectations to map. In this sense, the claim that acceptability matters to the design of AMAs is incredibly open ended, since the concept of acceptability itself can ostensibly cover very different evaluative paradigms. Accordingly, this chapter will explore the interplay between artificial morality (specifically top-down implementations of moral theory) and three plausible views of what constitutes acceptability: the moral preferences of society vis-à-vis the behavior of an AMA, the adoptability of an AMA (particularly in terms of the user's preferences for its behavior and functionality), and finally, acceptability understood as institutional viability, relating mainly to an AMA's adherence to what we have previously called *ethical design concerns*, such as the principles of transparency and accountability.

Since we have previously explored how such ethical design concerns can generally affect deterministic (or top-down) styles of programming in the first section of this thesis, our discussion of institutional viability will serve mainly as a framework within which we can assess the types of constraints all three senses of acceptability place on the design of artificial morality, through the resolution of what we will call *the problem of artificial moral uptake*. This problem, in turn, relates to the designer's identification of the morally relevant features (or marks of significance) in an AMA's *Umwelt*, or the process of identifying which types of facts or characteristics can ground whichever factors a moral theory holds to be normatively relevant (i.e. welfare or harm). This problem is particularly pertinent for the design of artificial morality, since it is likely here that the tension between, on the one hand, machine ethics understood as the imposition of moral constraints on an amoral machine, and on the other, machine ethics understood as the design of a machine which reasons or acts like a moral agent, reaches its most trying point. Put more simply, the problem of artificial moral uptake exposes the incompatibility between the internal constraints moral theory imposes on a robot's reasoning, and the external constraints that the moral expectations of society impose on that same process.

With these ideas in place, the chapter is laid out in the following way. In the first section, we explore the concept of acceptability as moral preference, leaning heavily on the results and methodology of MIT's Moral Machine Experiment. We will find that investigation into this corner of acceptability affords us valuable information as to the particularity of societal moral preference, as well as what we will call the *shape* of local common-sense morality. In section two, we delve into the notion of acceptability as adoptability. This will lead us to explore the concepts of behavioral equivalency and optimality in machine behavior, as well as the permissibility of special obligations towards an artificial moral agent's principal user. Finally, in section III, we address the complicated idea of acceptability as institutional viability, through the lens of what we will come to call the *problem of artificial moral uptake*. This will lead us to identify two problems—the problem of *perverse incentives* and the *further feature* problem—which significantly curtail the decisional efficiency of artificial morality.

1. *Acceptability as Moral Preference & the Scope of Artificial Morality*

Given the somewhat alarmist nature of recent avowals of the need for AI alignment, especially those which address the 'existential threat' these technologies pose to mankind as a whole[7], it seems plausible that many conceptions of artificial morality have taken on a distinctly *universal* tone in recent years. This is to say that given the ubiquity of technological artefacts which can be seen to operate in contexts of ethical salience, much recent work in machine ethics, especially proposals for concrete approaches to artificial morality, do not seek to provide a *domain specific* account of how moral responsiveness ought to be achieved, but rather address ethical responsiveness in artificial moral agents *generally*[8]. In a certain sense, this will to provide a universal blueprint for the structure of artificial morality reveals an espousal of what we have

[7] In effect, it would seem that recent breakthroughs in computing technology have led many prominent scientific figures to pronounce rather alarmist statements concerning the potential threat advanced AI poses for human society. Most notable among these is surely the late Stephen Hawking's claim that "The development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014).

[8] In the extensive implementation survey accomplished by Tolmeijer et al., they observe that "Most authors use a general approach to machine ethics: almost three out of four do not use a domain-specific approach, but focus on a general proposal of implementing machine ethics..."(2020, 19).

called the 'simple thesis'[9], where the focus has been on *how* to simulate moral reasoning in explicit ethical agents, rather than *what* the purpose of that moral reasoning is, or what it is meant to accomplish in a specific *Umwelt*.

One unfortunate consequence of this assumption, as we have seen, is the denial that different normatively relevant factors (hailing from different moral theories) can affect the computational decision procedure of an AMA's agent program. Another, broader misfortune, however is that this assumption denies the possibility that *moral relativism* may affect the shape and structure of artificial morality, or more loosely, that variance in the prevalent moral attitudes of different societies or sectors may not only affect *what* to reason about, but also *how* to reason. This assumption appears particularly short-sighted given the widely held conviction that moral relativism, or value pluralism, is not only a design constraint with which machine ethics must contend[10], but that the discovery of so-called universal, 'ground-truth' moral principles[11] cannot (or will not) arise in time to provide a universal *normative* framework for artificial morality[12].

Tangentially, these claims then seem to spell trouble for the maximalist, since if moral realism is false, then no single conception of the good may serve to inform the 'correct' or 'best' events for which artificial moral agents could be seen to be casually responsible. In other words, maximalists are at a loss for what to maximize. However, nothing prevents the maximalist from asserting that moral realism, despite appearances, *is* true, and accordingly, that observed variances in moral attitudes are nothing more than stubborn forms of moral parochialism which ought to be weeded out via the behavior of AMAs. As we shall see in the next chapter, this is precisely their claim, but we will leave it to one side for the time being.

What imports us here is the looser, pragmatic idea that since all of mankind at least appears to disagree about what matters morally, an artificial morality of a truly *universal* scope appears untenable for many machine ethicists. We are then left to ponder the appropriate scope of a given

[9] Gabriel, 2020.
[10] Gabriel, 2020: Himmelreich, 2018: Keeling, 2020: Evans et al., 2020.
[11] Noothigatu et al., 2018.
[12] Although some seem hopeful that a Superintelligent agent may one day access these normative truths (Bostrom, 2003: Yudkowsky, 2004).

account of artificial morality, or just how far a given normative framework could generalize across different societies, or different types of AMAs. It is precisely here where the most common conception of acceptability —but not the *only* conception—enters into play, as a tool to discover the contours of value pluralism. This entails the view of acceptability as an expression of *societal moral preference* and is surely championed by MIT's Moral Machine Experiment on autonomous vehicles[13].

This international study on the ethics of autonomous vehicle decision-making in dilemma scenarios has garnered an impressive amount of participation and acclaim, discovering a roughly global consensus surrounding certain action-guiding recommendations—spare humans over animals, spare more lives over less, spare the young over the old[14]—as well as a type of regional divergence concerning the *strength* of these recommendations—where, for instance, 'Eastern' countries were seen to exhibit a weaker preference for youth over age, and 'western' countries a stronger preference for saving the many over the few[15]. It would then seem that value pluralism and variance in moral attitudes can be given a geographical instantiation, which is to say, different approaches to artificial morality in autonomous vehicles may be appropriate for different regions of the world. In this sense, the *geographical scope* of a given type of artificial morality can be delineated.

However, while revelations concerning geographical variance in *what* can be seen to matter morally have been met with much acclaim, the study also shed light on another, more troubling, type of divergence in moral attitudes—the so-called *social dilemma* of autonomous vehicles[16]. This dilemma, the authors claimed, resulted from a type of incongruence across the expressed moral attitudes of participants, arising when participants moved from an *impartial* perspective—a detached view wherein the participant judged what ought to occur *sub specie aeternitatis*—to a more partial perspective, when the same participant was asked whether he was likely to purchase vehicles which were programmed in ways which might result in his death. Put more succinctly, Bonnefon et al. observed that judgement from the impartial perspective yielded a societal

[13] Bonnefon et al., 2016 : Awad et al., 2018.
[14] Dizikes, 2018.
[15] Awad et al., 2018.
[16] Bonnefon et al., 2016.

preference for "utilitarian cars"[17], those that spared the many over the few, while personal perspectives appeared to advocate for more egoistic vehicles, or those that would 'protect the passenger at all costs'. For Bonnefon et al., this discrepancy was seen to bear "…the classic signature of a social dilemma, in which everyone has a temptation to free-ride instead of adopting behavior that would lead to the best global outcome"[18]. Let us unpack this claim.

Essentially, it would appear that what steers Bonnefon et al. towards the use of the term 'social dilemma' is the idea of participants 'gaming the system' in ways which favor their survival and those of their loved ones (as passengers) in accident scenarios. This aligns with the idea of certain selfish members of the traffic community 'free-riding' off of the relatively cooperative preferences of others. In this sense, participants are seen to have two types of moral preference: first, an *internal*[19] preference for *their* vehicles to be programmed egoistically (in ways that favor their survival and those of their families), and an *external*[20] preference for the vehicles *of others* to be programmed in ways which may tip the scales in favor of the survival of *non*-passengers, in this sense maximizing the chances that this same participant (and his entourage) would never be sacrificed, regardless of their position in the traffic environment. Thus far, there is no inconsistency between these two preferences, since both point to individual (or familial) self-preservation. The trouble arises when Bonnefon et al. make the distinctly maximalist claim that such self-interested behavior fails to lead to the 'best' or 'correct' event, which they claim consists in a naive form of act utilitarianism which would minimize the total number of casualties occurring in the traffic environment. Indeed, in later analyses of Bonnefon et al.'s results, much attention has been payed

[17] This epithet, however, is misleading. In effect, the right-making feature of 'sparing the many over the view' is hardly reserved for expressly *utilitarian* forms of moral theory, since multiple moral theories can be seen to make this recommendation under certain circumstances. For example, Scanlon attempts to accommodate this feature in his espousal of contractualism (Scanlon, 1998, 229-241). We will nevertheless follow suit with Bonnefon et al.'s choice of vocabulary across our arguments for the sake of clarity.
[18] Bonnefon et al., 2016, 1575.
[19] Chauvier, 2013.
[20] Dworkin, 2013. "…the preferences of an individual for the consequences of a particular policy may be seen to reflect…either a *personal* preference for his own enjoyment of some goods or opportunities, or an *external* preference for the assignment of goods and opportunities to others…" (Dworkin, 2013, 234). Dworkin goes on to maintain that only personal (or what we call ''internal') preferences ought to be counted in a utilitarian calculus, since if this is not the case, "…the egalitarian character of [the] argument is corrupted, because the chance that anyone's preferences have to succeed will then depend, not only on the demands that the personal preferences of others make on scarce resources, but on the respect or affection they have for him or for his way of life" (Ibid., p. 235).

to these so-called 'biases' or shifts in moral attitude, occurring when individual participants adopt different evaluative perspectives[21]. As one such account in Frank et al. claims:

> The inherent problem of peoples' preferences in moral dilemmas, as discussed by Bonnefon and colleagues, is that people seem to favor a utilitarian moral doctrine that minimizes the total casualties in potentially fatal accidents, but they simultaneously report preferring an autonomous vehicle that is preprogrammed to protect themselves and their families over the lives of others. These findings illustrate that moral decisions could be a matter of personal perspective: When people think about the outcomes of the dilemmas for the greater good of society, they appear to employ a utilitarian moral doctrine; however, when they consider themselves and their loved ones, they shift towards a deontological moral doctrine that rejects the idea of sacrificing the passengers in their vehicle. As a consequence, moral codes derived from human decisions could reflect biased moral preferences[22].

Often, the use of the term 'bias' to describe these deontological forms of moral attitude reflects a commitment to the 'dual-process theory'[23]of moral reasoning, popular in moral psychology research. Very roughly, this theory points to a distinction between a deliberative mode of thinking—wherein individuals mobilize high degrees of cognitive resources—and an *intuitive* mode, where hard and fast decision-making is driven by emotions and 'easily accessible rules'. In a series of studies, Greene and colleagues find that deliberative modes of thinking yield more utilitarian moral decisions, while intuitive thinking yields deontological decisions[24]. Depending on the amount of temporal resources available at the time of decision, human beings can be seen to switch between these two modes[25].

It would then seem that the investigation of acceptability as moral preference, at least in the case of Bonnefon and colleagues, has in some sense served to bolster the validity of the dual-process theory of moral reasoning. In other words, we can observe that participants have made use of both modes of thinking, and that the shift between them appears to track the shift between

[21] Shariff, Bonnefon & Rahwan, 2017: Meder et al., 2019.
[22] Frank et al., 2019, 1.
[23] Kahneman, 2011: Evans & Curtis-Holmes, 2005: Epstein & Pacini, 1999.
[24] Greene et al., 2001.
[25] Greene et al., 2008.

personal and impartial perspectives in the evaluation of dilemma cases. Yet, maintaining that these two perspectives exist in the moral attitudes of society is a descriptive claim. Maintaining that either of the two is 'better', or that one ought to subsume the other, however, is surely a *normative claim* about the correct way to reason about dilemma cases in vehicle accidents. In other words, the sophistication of the evaluative mode of thinking, and the utilitarian-inspired decisions it generates, does not alone spell the moral 'rightness' of its conclusions. Acceptability as moral preference can tell us much about what matters morally for a given population, however it cannot tell us what ought to matter morally in the abstract. If a revision from agent-relative principles towards a more neutral concern for others is needed, it must be defended on normative, rather than descriptive grounds.

Mainly though, the association Frank and colleagues make between these two modes of thinking and the schools of deontology and utilitarianism lacks analytical clarity. Indeed, there is nothing *exclusively* 'deontological' about considering special ties and obligations as a normatively relevant factor of autonomous vehicle decision-making, save perhaps for the fact that these special ties themselves are non-consequentialist in nature. Similar to Bonnefon et al.'s use of the term 'utilitarianism' to describe the feature 'minimizing casualties', the activity of identifying societal moral preference is surely more profound than the activity of linking these preferences to naive forms of rival moral theories; a move which leads to the idea that these preferences are inextricably at odds, and thus that a hard choice must be made in the design of the vehicle's artificial morality. In this sense, an alternative reading of the social dilemma of autonomous vehicles may resemble the following: if it can be seen that *multiple normative factors* are seen to matter morally to a given population—in this case, a) an impartial concern for utility, b) a commitment to self-preservation of the passenger, and c) certain types of special obligations to the loved ones of the passenger or owner—then what acceptability research actually reveals is a moral attitude which, when taken as a moral theory, resembles an impure form of act consequentialism which has a pluralist conception of value, abiding by certain deontological constraints, or admitting certain special obligations.

This type of structure only poses a problem if the designers of the vehicle's artificial morality have antecedently committed to pure forms of utilitarianism, or for that matter, pure forms of deontological theory. Put another way, the fact that the moral attitudes of a population do not,

at the normative level, reflect a perfect instantiation of a pre-existing moral theory is only problematic if we have already decided which moral theory is best. Or, worse still, have set about designing a computational decision procedure which strictly abides by this 'best' or favorite 'theory' in terms of its decisional structure, building for instance, a structurally utilitarian autonomous vehicle which cannot cope with special obligations. In brief, an assumption of the 'simple thesis' may prove damaging here. But barring the technological impossibility of implementing these types of pluralist structures—which in this specific case, seems achievable given the 'cost and constraint' model of Gerdes & Thornton as discussed in the previous chapter—there does not seem to be any non-normative reason to improve upon the moral attitudes discovered by the Moral Machine Experiment, and thus the so-called social dilemma of autonomous vehicles loses at least some of its initial traction.

Regrouping these ideas together, it should be plain that the concept of acceptability as moral preference, especially when it is mobilized in empirical research, yields two important recommendations concerning artificial morality. Firstly, acceptability often reveals the limits of the *scope* of a given artificial morality, or to what degree a given moral theory or normative structure is conducive to the moral attitudes of a given society, and therefore particular rather than universal. In one sense, this yields a sort of 'best fit' recommendation for the application of a given moral theory to a population, where some populations are seen to be more 'utilitarian' than others, or to care more about the young than others, even if the cogency of this recommendation depends heavily on the data scientist's knowledge of moral philosophy. Metaphorically, we might then imagine the role of the designer of an AMA as something of a cosmic treasure hunter, waving his 'utilitarian', 'Rawlsian', or 'Smithian' metal detector across a map of the Earth, listening for that area in which it beeps the loudest. This, to various degrees, is what occurs when designers of top-down expert systems consider acceptability in their choice of moral theory, thus avoiding the strong maximalist claim that all moral attitudes can be incorrect, and for this reason should not be considered in the design of artificial morality[26].

[26] As Henry Sidgwick lyrically maintains, "…it seems that when we abandon the firm ground of actual society we have an illimitable cloud land surrounding us on all sides, in which we may construct any variety of pattern states; but no definite ideal to which the actual undeniably approximates, as the straight lines and circles of the actual physical world approximate to those of scientific geometry" (Sidgwick, 2019, 21).

However, another, more subtle way in which acceptability informs artificial morality, is through the revelation of what we might call the *shape of local common-sense morality*, even if, as we have seen, it is only when we abandon all theoretical commitments that we can see the value of these pre-theoretical intuitions. In this second sense then, acceptability shows us the shape of moral reasoning endemic to a given corner of the world, or more precisely still, a given *activity* in a given corner of the world. This is so since, while the Moral Machine Experiment may have shown that the shape of common-sense morality in western society roughly resembles a pluralist form of act consequentialism, it only did so within the context of dilemma decision-making in autonomous vehicles. In other words, if the Moral Machine Experiment has shown that common sense morality abides by this structure when addressing real-life car crash dilemmas, it certainly does not provide any evidence that this same structure holds, say, in the medical field generally, or in the dilemma decision-making of a robotic healthcare assistant. For this reason, the term *local* is important; pointing both to geographically-relative and role-relative moral attitudes, and also, to a specific *place* of moral reasoning, or a specific view—precisely, a *narrow* vision—of what can count as an ethically salient context[27]. In this sense then, the shape of common-sense morality that acceptability reveals may be highly valuable or insightful, but it is also quite limited. Consequently, it behooves the acceptability-conscious designer to carefully assess the data acceptability can provide, if he is to avoid mistaking trash for treasure.

2. *Give the People What They Want: Acceptability as Adoptability*

With moral concern surrounding the autonomy and implications of artificial moral agents reaching what is likely an all-time high, important and long-standing practical considerations can occasionally be pushed to the wayside. To this end, perhaps the principal oversight of the lion's share of recent literature on artificial morality is the idea that artificial moral agents are designed

[27] This claim is echoed by Winfield (2019), albeit in the context of bottom-up approaches to artificial morality. As he sees it, when using empirical data derive or 'learn' decisional principles, "…we need to be absolutely certain that the dataset has not been biased and that the explanatory principle so learned really does have predictive leverage when applied to a different context. For sure, if a machine learns a 'wrong' or 'inadequate' principle, or even just a 'simple' principle, then there will be problems if we try to apply it in other situations" (2019, 513).

with a purpose-oriented ontology; current level 4 explicit ethical agent AMAs are meant to accomplish something *practical* in the human social sphere, and it is this practical purpose that leads an individual or institution to employ these agents in the pursuit of their practical ends. This rather 'tool AI' notion is sometimes lost in the larger debates on AI alignment, mostly because the object of analysis is either a) an imagined *universal* practical agent or AGI, or b) the ubiquitous existence of automation considered generally, where both imply multiple practical purposes (and multiple *Umwelts*).

One familiar, albeit comical upshot of this claim is the idea that no (current) artificial moral agent is designed to act as a *pure moral saint* in the human social sphere[28]. This is to say that the role of a given AMA will always be tethered to some practical purpose, or the achievement of some practical end. As we maintained in the previous chapter, we have not, and likely will not, in other words, design a Philanthropotron3000™, whose extensional restrictions amount to 'doing good in the world', abstractly considered. It follows then, that the successful achievement of an AMA's practical purpose plays a significant role in an AMA's acceptability, an aspect which we can capture with the notion of *acceptability* as *adoptability*, or as it is sometimes called in human-machine interaction literature, technological *acceptance*[29].

Given our discussion so far, the apparent importance of acceptability as adoptability would seem to indicate two sets of normative standards in the evaluation of AMA behavior. The first set, familiarly, pertains to the *moral expectations* human agents hold vis-à-vis the behavior of an AMA, or the types of moral principles, values, or claims that their behavior is expected to be responsive to, and perhaps the appropriate degree of this responsiveness. In this sense, acceptability as moral preference amounts to a particular (individual or collective) judgement about which of these elements is appropriate for inclusion in the artificial morality of an AMA which acts in a certain context, or in a certain capacity. We expect, in other words, that autonomous vehicles minimize the number of injuries caused in vehicle accidents, but also expect this same vehicle to exhibit a certain allegiance to its passenger and his welfare[30]. As we have seen, these expectations are

[28] Grau, 2006.
[29] Davis, 1989 : de Graaf et al., 2019.
[30] Lin, 2016 : Keeling et al., 2019 : Evans et al., 2020.

derived from the moral attitudes a particular society holds—and are thus limited in *scope*—even if some more abstract attitudes or prescriptions, such as 'a robot should not cause harm to humans', could plausibly hold across many, if not all *Umwelts*[31].

The second type of normative standard pertains to what might be described as *functional expectations*. As the name implies, at a general level, this standard tracks the expected *usefulness* of an AMA in the pursuit of a given end, where the AMA is conceived as a tool for human agency. Even if we consider the world of *mundane* technical artefacts, it is not difficult to ascertain the importance, and in some ways, the *power* functional expectations have over our use of technology. For instance, if my aim is to clean a very dirty floor in a modern household, a number of means are likely at my disposal towards this end. Ostensibly, I could a) use a broom and dustpan, b) use an electric vacuum, or c) my hands. There is of course an obvious betterness relation which holds across these methods, which relates to the efficiency of each: a broom is more efficient than my hands, and the vacuum is more efficient than the broom. Accordingly, if I elect to use a vacuum, my expectation is that it allows me to clean the floor better than I could by either of the two alternatives. My use of the vacuum is therefore dictated by my perception of its relative optimality, or put another way, my adoption of the tool is conditional on its being the most efficient available option to achieve my end[32]. In this sense, I would certainly experience disappointment, if not frustration, if it turned out that the vacuum I elected to use *failed* to meet my expectations for efficiency; if for example the vacuum had no suction, preventing it from removing anything larger than a grain of rice from the floor.

In a rough sense then, functional expectations take the form of an evaluative threshold; one which provides two types of standards which may dictate the adoptability of a technology. First, a

[31] Nevertheless, whether or not this preference is satisfiable will often depend on the context of implementation—perhaps autonomous vehicles, and certainly autonomous weapons, would seem to point to this impossibility.

[32] This is perhaps an oversimplified version of the Technology Acceptance Model (TAM), popular in human-robot interaction research (Davis, 1989). Under this theory, the acceptance of a technology is dependent on the perceived usefulness of the technology, and its perceived ease of use, where a user can form intentions—and eventually attitudes—towards the use of a technology along these variables. While critics of this model find it too rudimentary for application to the acceptance of complex technologies such as social robots (de Graaf et al., 2019: Van den Poel, 2016) for our purposes here, it nevertheless neatly outlines the basis of what we have called a functional expectation.

minimal threshold which points to the idea that a machine must perform 'as good as' a human agent in its task or role for it to be useful to an individual. To use our previous example, a vacuum fails to meet this minimal standard when it proves less efficient than my hands at cleaning the dirty floor. We might reformulate this idea through what we can call the *principle of behavioral equivalency*:

The Principle of behavioral equivalency (PBE): for a technology to be adoptable, it must perform at least as well as a human agent in the execution of its function, role or task.

At first blush, the PBE may seem to reflect not only the very reasonable expectations of the user or larger society, but perhaps the very utility a particular designer or original equipment manufacturer finds in investing in the development of a technology in the first place. This is to say that it does not make much economic sense to build machines that hinder individual human agency or hamper collective action through suboptimal performance. Indeed, it is even reasonable to assume that technologies which fail to attain behavioral equivalency likely will not make it to market. However, this pure interpretation of the PBE is likely frustrated by some of the relatively unique effects that artificial (moral) agents can have on socio-technical systems. To this end, there are at least three considerations which could complicate a simple view of the principle of behavioral equivalency.

First, there is the relatively straight-forward consideration of the benefit of automation—regardless of its quality—to the efficiency of a user's practical agency. In this sense, users may nevertheless use and adopt robot vacuums, even if they fail to clean a floor as well as a human agent would. Most likely, an individual's choice to adopt this behaviorally sub-optimal technology will revolve around a) the consideration of the opportunity cost of cleaning the floor one's self, or b) the cost efficiency of purchasing a robot over employing a maid to perform the same task.

Secondly, and more central to social robotics, there is what is often called the 'Eliza Effect', which points to "the susceptibility of people to read far more understanding than is warranted into

strings of symbols—especially words—strung together by computers"[33]. In this sense, a human user may project or attribute emotions, intentions or 'intrinsic qualities and abilities' to machines, which far surpass their actual technical capacity—such as when a user mistakes Siri's deterministic albeit comical responses to questions such as 'do you love me?' as an instance of a veritable and original sense of humor[34]. Sherry Turkle, in a similar vein, calls this the 'as if' self, where a robot behaves 'as if' it had emotions, gratitude or meaningful bonds of friendship, thereby tricking its user into engaging in misplaced and cognitively dissonant forms of relations with the technology[35]. In this sense, even if a chatbot does not engage in conversation 'as well as' a human agent does, human users may nevertheless use and adopt the technology, either because a) the user projects subjective qualities and capacities onto the chatbot which help close this gap, or b) because the user willingly and even knowingly behaves 'as if' this chatbot were behaviorally equivalent.

Finally, and in a somewhat connected sense, there are a host of ways in which sub-optimal behavior and automation can be accommodated and adopted by human agents, all of which revolve around the dissolution of what is called the 'substitution myth' of autonomous technology: "as machines acquire more autonomy, they will work as simple substitutes (or multipliers) of human capability"[36]. In effect, the introduction of automation into cooperative environments may not only shift the degree, scope, and responsibilities of task-sharing required to achieve posited goals, but may even change the nature of the tasks each actor is meant to perform, often requiring an entirely different configuration of human skills[37]. For instance, the inclusion of automated surveillance and detection systems in military contexts often requires humans to work faster, do more, or perform their roles in more complex ways—what is often called the *law of stretched systems*[38]—or, increased automated capacities require increased supervision and surveillance on the part of human agents, which leads to non-negligible and continuous operation costs[39]. In this sense, while the individual artificial agent may not be behaviorally sub-optimal when taken as an individual actor,

[33] Hofstadter, 1996, ix.
[34] For English versions of this program, the typical response is "Let's just be friends".
[35] Turkle, 2008, 313-315: 2017.
[36] Bradshaw et al., 2013, 60.
[37] Ibid, p. 60: Christofferson & Woods, 2002: Norman, 1991.
[38] Bradshaw et al., 2013: Adams, 2001.
[39] Bradshaw et al. 2013, 61.

the impact its performance has on a socio-technical system as a whole may decrease the overall efficiency of cooperation in subtle ways.

In light of these complexities, it may appear that the attainment of behavioral equivalency in artificial agents already poses significant technical challenges, and further, that this principle is perhaps best applied to individual AMAs interacting with a single or 'principal' user, what we have called *surrogate agents* in previous chapters. However, it is certainly plausible that behavioral equivalency alone may not drive the robust adoption of a given technology. Rather, it would seem that many users, and likely the larger society to which they belong, expect that artificial agents perform their tasks *better than* a human agent could in similar circumstances. Indeed, this type of reasoning often provides the principal justification (and perhaps even moral justification) for the implementation of many types of artificial moral agents, particularly autonomous weapons and autonomous vehicles. We can capture this idea with a second principle or *maximal* threshold which likely reflects the functional expectations of human agents, what we will call the principle of behavioral optimality:

The principle of behavioral optimality (PBO): for a technology to be highly adoptable, it must perform better than a human agent in the execution of its function, role or task.

Using autonomous vehicles as an illustration, it would indeed appear that functional expectations towards an autonomous vehicle would abide by the principle of behavioral optimality. First, at the collective level, this seems accurate given the onslaught of media attention surrounding the purported end of vehicle-related casualties relating to human error: no more drunk, distracted, dangerous or road-raging drivers on the world's roads[40]. Indeed, some optimistic figures point to an estimated 90 percent reduction in the 1.35 million individual deaths caused by vehicle accidents each year[41], a reduction which is directly related to the behavioral optimality of autonomous vehicles. In a more individual sense, it seems that if I elect to use an autonomous vehicle to drive to the airport, say, it must be because the autonomous vehicle presents the best, most efficient, or most useful option to me; dominating the alternative options of either driving myself, taking a taxi,

[40] Lin, 2014b: 2017.
[41] Airbib & Seba, 2017: Fagnant & Kockelman, 2015: Gao, Kass, Mohr & Wee, 2016.

walking, or taking public transportation[42]. Of course, the precise meaning of 'efficiency' or 'usefulness' is somewhat vague here, since just as in the case of behavioral equivalency, many factors may influence my perception of usefulness: cost-efficiency, time, relative ease of access, comfort, etc. In other words, it may not be the case that for autonomous vehicles to be *adoptable*, they will need to be strictly superhuman *drivers* from a technical or strategic standpoint. It may suffice, for instance, that the benefits of being able to engage in on-board activities—or the desirability of being able to do things *other* than keep one's eyes on the road—be sufficiently attractive to ensure adoption. Thus, the *experiential* optimality, rather than the strictly behavioral optimality of an autonomous vehicle may be at work behind my perception of its usefulness.

Still, I would likely experience a very high degree of frustration if, during my trip to the airport, my autonomous vehicle exhibited any of the following behaviors: never exceeding the speed limit, even if neighboring vehicles are whizzing past me; always stopping at an exceedingly safe distance for every pedestrian, including unlawful j-walkers; never running yellow lights; never changing lanes unless the risk of oncoming traffic is infinitesimally small; or demanding that I take control of the vehicle when weather conditions reduce visibility, or worse still, when an unavoidable collision is imminent. Strangely, while most of these behavioral traits surely constitute a *safe* autonomous vehicle[43], they nevertheless point to a very 'mild-mannered' or risk-averse vehicle, one that likely does not drive how *I* would have driven in many of these contexts. In this sense, even if the vehicle is safer, or certainly more law-abiding, I may nevertheless find it behaviorally or experientially sub-optimal, and thus relatively unadoptable. There must be a certain degree of concordance between my functional expectations (or preferences) and the outward functional behavior of the vehicle. Put another way, how the vehicle drives, and how I would have driven, must be sufficiently similar for me to adopt the technology, or find it useful[44]. For the vehicle to be easily adoptable, its behavior and mine should be functionally equivalent.

---

[42] This assumes the slightly more European model of the prospective form of implementation of autonomous vehicles, one in which they exist as a public service, or 'robotaxi'. A more North American model, one which assumes private ownership, can nevertheless be accommodated if we imagine the driver choosing between modes of autonomy, i.e. driving himself, or turning on 'autopilot'.

[43] This is perhaps an oversimplification, since in many cases—especially in terms of adherence to speed limits— exceedingly lawful behavior can generate increased risks for the general traffic environment.

[44] In the media, this concern for the experiential satisfaction of the passenger derives as much from the internal behavior and decisions of the machine, as it does the external behavior of *other* road users in response to the presence of 'mild-mannered' autonomous vehicles (Condliffe, 2016). One major concern

In this sense, the boundaries between behavioral optimality and equivalency are blurred by the changing perspectives of adoptability. As the case of autonomous vehicles makes plain, the behavioral optimality of 'safe' autonomous vehicles may be challenged by the will to achieve behavioral equivalency in terms of the individual passenger's preferences for his vehicle. Thus, the barriers to adoption, understood as the satisfaction of functional expectations, may not neatly align in the design of an artificial moral agent. To this end, and taking these aspects together, it would appear that functional expectations provide two types of constraint on the design of artificial moral agents. The first, more general constraint pertains to the extensional restrictions of the AMA: for a human agent to employ an AMA in the pursuit of his ends, his functional expectations must match the functional purpose of the machine—BerryPicker3000™ must pick berries, and an autonomous vehicle must drive autonomously from one desired location to another. As soon as the AMA fails to meet this first type of functional expectation—or perhaps, achieves ends *other* than those the user expects—the machine may no longer be a viable means for the user's end, and in this sense its acceptability as adoptability is threatened. Pragmatically however, this constraint, understood broadly at least, is easily met by the brunt of current AMAs, so long as their users are suitably informed as to the specifics of the machine's functionality[45].

A more complex design constraint lies in the effects functional expectations have on the *intentional restrictions* of an AMA, which we will recall, provide the 'strategy' or decision procedure by which an AMA deliberates upon and selects its actions, and thus achieves its goals. Here, functional expectations appear to provide a sort of qualitative threshold for intentional restrictions: on the one hand, for a machine to be highly adoptable, it must be behaviorally or experientially superior to any other comparable means by which to achieve its particular purpose.

---

lies in the possibility that risk-averse autonomous vehicles will generate a traffic environment in which pedestrians will be able to act with impunity, to the chagrin of most passengers (Millard-Ball, 2018).

[45] Our discussion of adoptability shares much ground with a parallel discussion on the concept of user trust, where trust is often considered to be a necessary condition for the adoption of an intelligent artefact. For instance, the idea of concordance between a user's functional expectations and the machine's extensional restrictions is rather neatly mirrored by Mark Coeckelbergh's concept of 'trust as reliance', where "…we expect the artefact to function, that is, to do what it is meant to do as an instrument to attain goals set by humans. Although we do not have full epistemic certainty that the instrument *will* actually function, we expect it to do so. For example, one may trust a cleaning robot to do what it is supposed to do—cleaning" (Coeckelbergh, 2012, 54).

This does not imply that the machine must perform better than a human strictly speaking, however it does imply that adoptability is at least loosely correlated to efficiency, understood in this broader sense. On the other hand, functional expectations also appear to provide something of a negative threshold, one which prohibits certain behavioral traits or intentions, whenever they diverge significantly from the expectations, preferences or intentions of the machine's principal user. The wider the divergence, the less adoptable the machine becomes.

In the literature, the influence and necessity of this latter type of constraint has been widely discussed via the larger theme of user participation[46], delegation, or involvement[47]. Interestingly, while in the field of human-machine interaction it is somewhat generally accepted that user participation is at least desirable in the design of AMAs[48], the machine ethics community has adopted a decidedly moralistic stance on the issue. One particularly virulent example of this can be seen in the debate surrounding the possibility of *mandatory ethics settings* in autonomous vehicles[49]. While many authors concede that user input is, to various degrees, a significant condition of adoptability, the moral cogency of allowing adjustable ethics settings seems to involve the weighing of two very different types of argument. On one hand, some authors have pointed to the inherently liberty-limiting repercussions that the enforcement of mandatory ethics settings would surely bring about, focusing on either a) the damage this type of policy would cause to the principles and values of individual autonomy[50], or b) the more general idea that such an imposition appears to thwart the principle of axiological neutrality which is inherent to the moral structure of many liberal societies[51]. Together, these concerns seem to advocate for the idea that an individual user may have a political, if not a moral right to select the principles by which her vehicle makes decisions, especially if these decisions involve significant threats to her welfare, or life-or-death scenarios. As Jason Millar maintains, this freedom to choose flows from a conception of an

[46] Dignum, 2019.
[47] Tolmeijer et al., 2020.
[48] Etzioni & Etzioni, 2017: Tavani, 2015: Johannsen, 2009.
[49] Gogoll & Mueller, 2017: Contissa, Lagioia & Sartor, 2017: Sandberg & Bradshaw, 2013: Millar, 2014a: 2014b: 2015: 2017: Lin, 2016.
[50] Sandberg & Bradshaw, 2013: Lin, 2016.
[51] Gogoll & Mueller, 2017: Millar, 2014b: 2017.

autonomous vehicle as a moral *proxy* for its passenger[52], rather than an (artificial) moral agent or patient, *stricto sensu*.

Arguments which seek to defend the (moral) necessity of mandatory ethics settings, on the other hand, can be seen to revolve around two themes: first and mainly, that the affordance of adjustable ethics settings would lead to a prisoner's dilemma or the 'crowding out of morality' over prolonged strategic interaction and iteration, since many individuals will not be inclined to select more altruistic ethics settings, and thus 'defect', if it can be seen that most of their fellow commuters have opted for more egoistic settings[53]. Since egoistic ethics settings can be seen to lead to a world where total casualties are 'necessarily higher', these authors deduce that "…. selfish as well as moral agents have a strong reason against implementing [personal ethics settings]"[54]. A second central claim, which relies on more maximalist intuitions, is that personal ethics settings will lead to a world of biased and morally troubling decisions, since as Patrick Lin sees it, "…saving, protecting, or valuing one kind of thing effectively means choosing another kind to target in an unavoidable crash scenario"[55].

While such a world is certainly ostensibly undesirable, the cogency of this latter argument clearly hangs on the criteria by which the winners or losers are chosen in this supposed zero-sum game. Under one particularly alarmist reading, this discrimination is made along morally dubious personal characteristics such as age, gender, or sexual orientation[56], following in the footsteps of some of the variables chosen by the Moral Machine Experiment. In Lin's actual interpretation however, and picking up on a thought experiment introduced by Noah Goodall[57], the characteristics in question resemble more inoffensive features such as 'helmet wearing' or 'non-helmet wearing cyclist'. Nevertheless, both accounts of this problem miss the mark in two ways. First, if it is possible that personal or adjustable ethics settings are morally permissible to

[52] Millar, 2017: Keeling et al., 2019: Evans et al., 2020.
[53] Gogoll & Mueller, 2017.
[54] Ibid., p. 14.
[55] Lin, 2014a.
[56]"… [Personal Ethics Settings] might allow options that seem morally troubling: for instance, targeting black people over white people, poor people over rich ones, and gay people over straight…" (Gogoll & Mueller, 2017, 8).
[57] Goodall, 2014: 2019.

implement, it certainly does not follow that the user necessarily has an *unbridled* authority over who lives and dies, and along what criteria. Put another way, the moral permissibility of personal ethics settings does not logically point to the moral permissibility of just *any* ethics setting, so long as it is devised by the (morally bankrupt) passenger[58]. Further still, a passenger might not require such an extensive smorgasbord of personal criteria in order to give voice to his preferences, and indeed, such an extensive scope of choice may exacerbate the problems of *moral overload* endemic to the practice of user participation[59]. In this sense, beyond any *moral* reasons we might have to restrict choice options to more savory or impersonal criteria, there likely exists other practical and legal reasons to keep things simple[60].

Secondly, allowing such a fine-tuned degree of user participation appears to presuppose a type of 'traffic community prioritarianism', where the more a given individual satisfies the preferences of a user, the more likely he is to survive a lethal collision with that user's autonomous vehicle. Nothing in the concept of a personal ethics setting seems to mandate this particular approach to moral reasoning, and indeed, in related work investigating the moral permissibility of 'ethical dials'[61], other less alarming configurations are proposed, including a dial which modulates the value trade-offs of mobility and safety, and the interests of the passenger versus the external environment[62]. In this sense then, even if the intuitive attractiveness of personal ethics settings gains much of its impetus from the apparent value of incorporating a user's personal preferences

[58] A similar claim has been made by Etzioni & Etzioni in their defense of the virtues of what they call 'ethics bots', machines which learn and eventually emulate the moral preferences of their end-users: "One may ask: what if these preferences are harmful?…This question and similar ones do not take into account the major point we cannot stress enough: that the ethical decisions left to the individual are only those which the society ruled—rightly or wrongly—are not significantly harmful, and hence remain unconstrained by regulation or attendant legislation…ethics bots only address areas left open-ended by the law"(2017, 416-417). In the next section, we will further address how the use of these types of morally dubious features encounter significant resistance from adoptability as institutional viability.

[59] Van den Hoven, Lokhorst & Van de Poel, 2012. "The basic idea of moral overload is that an agent is confronted with a choice situation in which different obligations apply but in which it is not possible to fulfil all these obligations simultaneously" (2012, 144). In one sense, moral overload describes the situation of the engineer of an intelligent artefact when faced with tough value trade-offs such as that of safety and efficiency in design. In other sense, it describes the situation of an end-user who must make these value-laden choices in the form of customizable ethics settings, choosing for instance, more egoistic or altruistic collision algorithms.

[60] Lin, 2016: Gurney, 2015.

[61] Contissa, Lagioia & Sartor, 2017.

[62] J. Himmelreich in (Evans, 2019).

into the decisions of his vehicle, it does not follow that the personal *characteristics* of others are the only way he can express his moral attitudes.

Thus, with these concerns mitigated, we are left to contend with the by now familiar claim that a world filled with passenger-protecting autonomous vehicles—in this case achieved via adjustable ethics settings—fails to lead to the best outcome: the total minimization of causalities which could result from autonomous vehicle accidents. In the previous section, we outlined one way to diffuse this social dilemma, by pointing to the complexity of local common sense morality as it is revealed by empirical research into acceptability as moral preference, and thus to the idea that *both* impersonal utility and passenger-centric special obligations can be seen to matter morally. This led us to conclude that an artificial morality which could address both of these features may be both achievable and acceptable, rather than making the distinctively maximalist claim that one of these features ought to be subsumed by the other. Interestingly, the concept of acceptability as adoptability provides further support for this type of policy, and indeed, is likely the principal means by which one typically argues for the moral value of putting the passenger first.

Stepping back briefly, it would seem that taking up the perspective of acceptability as adoptability brings us close to a distinction made in chapter three, between what we have called a *surrogate* agent and a *distributive* agent. A surrogate agent, as we might recall, denotes an AMA whose agency serves as a proxy for a *particular human user*, while a *distributive agent* acts more 'autonomously', on behalf of no particular human agent. In terms of intentional restrictions, this difference tracks on one hand, the pursuit of a user's (idealized) interest in the AMAs practical agency, and on the other, the pursuit of the 'general' or 'collective' interest, often via the optimal performance of a given role or activity. Importantly, surrogate and distributive agents do not constitute deep ontological categories, rather, this distinction is phenomenological, made in the perception a human agent holds concerning the AMA, and by extension, in the intentions he forms as to its purpose and utility (for him). Returning now to the example of autonomous vehicles, it seems plausible that acceptability research has indicated a troubling type of *double vision*: passengers perceive their vehicles as surrogate agents, but society at large—and certainly some

machine ethicists—perceive autonomous vehicles as pure distributive agents, acting in pursuit of the collective interest. How salient is this distinction for the design of artificial morality?

Hypothetically, this distinction would prove salient indeed if surrogacy were a condition for the adoptability of autonomous vehicles. This is to say that if the reasonable pursuit of the passenger's (ideal) interest figured in the passenger's functional expectations, then it would appear not only that a) any pursuit of the collective interest in decision-making would be perceived as a case of machine *misconduct*[63], but also b) that many passengers would find these autonomous vehicles *unadoptable*. Here, we find the most damaging assumption made by the decision-theoretic appraisal of autonomous vehicle decision-making, and in a broader sense, the moral maximalist: namely, that optimality or maximization for the collective interest is impervious to the original incentives individuals have to adopt autonomous vehicles. It is not clear whether the pursuit of the collective interest at the level of an individual AV's *artificial morality*, will actually lead to the maximization of the collective interest of *society generally*, if people are disinclined to step into such a vehicle in the first place. In simpler terms, if every vehicle is programmed to minimize casualties, but people are disinclined to adopt vehicles programmed this way, then it is unclear how many casualties will actually be avoided by autonomous vehicles, since there may not be a sufficiently large number of them on the world's roads. In this sense, our issue is less with the larger *criterion of rightness* these views uphold—to minimize total casualties in vehicle accidents—and more with the specific individual *decision procedure* they view as the correct (or only) route to achieving it—implementing sacrificial or utilitarian forms of artificial morality into autonomous vehicles. Depending on the strength of a potential passenger's aversion to 'passenger-sacrificing' autonomous vehicles then, it may be the case that autonomous vehicles never get the chance to be 'better than' human drivers, even if we believe this is the best or correct outcome.

[63] This charge of misconduct is likely further substantiated by the trust relations the user holds, or expects to hold, with her vehicle. As Mark Coeckelbergh explains, "…trust ascription creates a deontic field: if someone trusts me, I feel under an obligation not to misuse that trust" (2011, 55). In this sense, if we can accept that a passenger may enter into such trust relations (regardless of whether these are appropriate given the ontological or moral status of the autonomous vehicle), then this likely provides further support for the normative relevance of special obligations between the passenger and her vehicle, and thus for their inclusion in an autonomous vehicle's artificial morality.

Generally then, ignoring the power of acceptability as adoptability can lead us to a very narrow vision of the purpose of artificial morality, one which, bluntly, often assumes that the technology is either already accepted, or does not need to be desirable or useful to exist. Indeed, in doing so, we are very often putting the cart before the horse, wondering how to optimize the moral behavior of an AMA which, in all likelihood, no one wants to use or purchase. When we fail to take adoptability seriously, our views of the types of moral principles, values and dispositions which are appropriate or necessary features of an AMA's artificial morality tend to approximate those which seem appropriate for Philanthropotron3000™, a pure distributive agent with no practical purpose other than to maximize the Good. Unfortunately, Philanthropotron3000™ is as useless to the user as it is unmarketable for the original equipment manufacturer, and in this sense, would seem to cast doubt on the viability of such impersonal principles in the real-world implementation of artificial morality.

The main upshot of an appraisal of acceptability as adoptability then, is the inclusion of user-centric or passenger-centric considerations in the general structure of an AMA's artificial morality. Indeed, adoptability allows us to argue their moral importance in two ways. One, *direct* type of argument holds that user-centric obligations may be morally justified under a conception of AMAs as moral proxies for their users, or in virtue of the trust relations users may hold with these machines[64]. In this sense, this direct view holds that a user's autonomy—and perhaps by extension, a respect for his interest—is a morally salient feature of many AMA *Umwelts*, and accordingly, that the structure of an AMA's artificial morality must accommodate this moral value, either via user participation, or through user-centric forms of intentional restrictions. This claim seems particularly cogent in situations where the user perceives the AMA as a surrogate agent, acting on behalf of its user. To be sure, many current and plausible types of AMA could easily be perceived as surrogate agents; most obviously, those which replace domestic or private activities which were hitherto accomplished by a human agent: robot vacuums, many social robots, virtual assistants, and perhaps autonomous vehicles. In all of these cases, the general justification for user-centric considerations flows from the idea of a user's relinquishing his autonomy or control: by allowing an autonomous vehicle to drive me from A to B, I relinquish all of the autonomy of means

[64] Millar, 2015: 2017: Keeling et al., 2019: Evans et al., 2020: Coeckelbergh, 2011.

I would have in deciding how to arrive at B. I might therefore be owed a certain allegiance[65] from the vehicle, which could take the form of a respect for my preferences.

Interestingly however, not all AMAs are easily conceived as surrogate agents. Indeed, autonomous weapons—either in the military or in domestic police forces—and to a certain extent, robotic medical assistants, do not seem to be obviously beholden to the preferences or interests of one particular individual[66]. Indeed, it seems that the pursuit of private interest—or the private interests of an institution or government—would be morally troubling in these agents: a robotic healthcare assistant likely fails to perform its role if it secures a kidney for a 'favorite' or 'celebrity' patient, when others are in more urgent need of a transplant. Nor would it be morally agreeable if police surveillance robots targeted a particular type of individual, based on personal characteristics such as race or country of origin. Instead, these types of AMA appear to require a very impartial type of artificial morality if they are to be acceptable from a moral standpoint, one that likely aligns with a particular conception of justice. The distinction between surrogate and distributive agents is then useful, since it affords further insight into the contextualization of an AMA's artificial morality; pointing to the proper purpose of artificial morality as serving either the individual or the collective interest, or in special limiting cases such as autonomous vehicles, a bit of both[67]. Without this distinction—which is often not made in the machine ethics literature—we are blind to the varieties of moral responsiveness which are likely appropriate for different types of AMAs, a shortsightedness which typically leads to the undermining of a user's autonomy and the design of purely distributive types of artificial morality. In short then, investigating adoptability may reveal the role of an AMA as a surrogate agent, and thus, provide normative grounds for including user-centric considerations in that AMA's artificial morality.

[65] Lin, 2016.
[66] We omit here the limiting case of robotic healthcare assistants which operate in a user's home, which much like autonomous vehicles, appear to hold a 'double status' of surrogacy and distribution.
[67] In our discussion in this section, we have tacitly assumed that an autonomous vehicle is either privately owned or 'hired' by a particular individual, resembling something of a robot taxi. This is forgivable given that the brunt of literature on autonomous vehicle ethics makes a similar assumption. However, an AV's double status as a surrogate and a distributive agent would likely shift if these vehicles were not privately owned, but instead constituted autonomous modes of public transportation, or emergency vehicles.

However, acceptability as adoptability also affords us an important *indirect* argument for the inclusion of user-centric considerations in artificial morality. This argument holds that if adoptability is a necessary condition for the ubiquity of a behaviorally optimal technology, and if adoptability hinges on the inclusion of user-centric considerations, then user-centric considerations must be included, at least *preliminarily*, if the benefits of behavioral optimality are to be secured. Here, the qualifier 'preliminarily' is important, since it points to the idea that the structure of an AMA's artificial morality need not be permanent or eternal. Indeed, even if the minimization of casualties in vehicle accidents is the correct or best event to bring about in the case of autonomous vehicles, it is likely the case that the artificial morality which best achieves this goal will shift over time: as people grow accustomed to the technology, and as the traffic community moves from predominantly human-driven cars, to a mixed fleet, and finally to a world in which AVs constitute the majority of vehicles on the world's roads. In this sense, conceiving of the need for user (or passenger) centric considerations as a *moral failure that must be immediately overcome*, rather than as a *temporary concession required for the attainment of optimal ends*, seems particularly short-sighted. There are likely limits, in other words, not only to how *much* an AMA needs to satisfy its user's (potentially immoral) preferences, but also for how *long* such a strong obligation towards the passenger is required for adoptability to be secured. Generally then, adoptability can indicate the initial 'moral price of entry' for a morally beneficial technology, but need not provide an atemporal recommendation for the structure of its artificial morality.

3. *Acceptability as Institutional Viability: The problem of Artificial Moral Uptake*

Until this point, our evaluation of the concept of acceptability has taken a somewhat democratic bent: acceptability as moral preference seeks to discover the influence that local common sense morality might have on the design and purpose of artificial morality, while acceptability as adoptability mainly seeks to establish the role and importance of the user (and his interest) in said pursuits. On both accounts, taking acceptability seriously loosely amounts to giving the people what they want in the design of artificial morality. There is a final, and relatively top-down conception of acceptability, however, which seeks to ensure that the artificial morality of an AMA abides by certain ethico-legal constraints laid down at the institutional level, in the

form of international policy recommendations, expert commissions, or the elaboration of state doctrines on the design of specific types of AMA. To keep things simple, we will call this interpretation *acceptability as institutional viability*.

Since this final form of acceptability constitutes something of a heterogenous category, including ethical design concerns, the question of legal liability and responsibility, and notions such as human or civic rights, it should not be surprising that institutional viability often seriously constrains the design and implementation of artificial morality. At its most invasive, institutional viability may supplant the need for artificial morality *entirely*, providing its own framework of norms and rules which provide action-guiding recommendations to the AMA's agent program in ethically salient contexts. In the field of autonomous vehicle ethics, for instance, this type of claim is often made by those authors of a more legal persuasion[68]. In a second, more moderate sense, a respect for institutional viability can, as we have seen across previous chapters, affect the design of an agent program, demanding for instance a top-down rather than bottom-up approach in a given AMA so as to bolster algorithmic transparency and predictability. There is however, a final and more subtle sense in which institutional viability can affect the design of artificial morality, one that provides an institutional answer to the question 'how much *should* a machine know?'. Arguably this final sense constitutes the thin—and often overlooked—edge of the intersection between acceptability and artificial morality, and accordingly, it will be our focus here.

In detail, the question of 'how much a machine should know' directly relates to the identification of morally relevant features (or marks of significance) in an artificial moral agent's *Umwelt*. In the previous chapter, we touched upon the idea that there is a necessary connection between the type of moral theory chosen for implementation on the one hand, and the types of facts or contextual features it holds as morally relevant, on the other; where the designer is left more or less interpretive leeway in identifying how these features are best instantiated by the environment. In the case of a welfarist form of utilitarianism, for instance, the only morally relevant feature is (expected or overall) *welfare.* Thus, in order to implement this type of theory, the designer must identify—and ideally, render the AMA sensitive to—every feature or fact about

[68] Casey, 2016.

a particular *Umwelt* which could contribute to, or be significant for, the assessment of the welfare of human agents[69]. Let us call this activity of identifying morally relevant features and their constitutive marks of significance *the problem of artificial moral uptake*.

Obviously, what can count as an indication of welfare—or for that matter, any morally relevant feature—may vary widely across different contexts: in the case of social robots, moods and emotional cues could signal dissatisfaction or approval, which might indicate an individual's welfare, where in autonomous vehicles, the expected degree of harm a pedestrian may incur as a result of a collision with the vehicle could certainly be indicative of his welfare. In this sense, the scope or limits of the AMA's *Umwelt* can help the designer in the identification of these features, specifically by providing an increasingly restricted account of what can instantiate a morally relevant feature. Importantly, however, if the resulting artificial morality is to be sufficiently robust, it does not suffice to identify the morally relevant features of one particular case, or one particular dilemma. When a designer attempts to solve the problem of artificial moral uptake, he must consider the AMA's action across the entirety of its *Umwelt*, an assessment which necessarily covers many particular moral cases (or configurations of marks of significance) to which the AMA must respond[70]. For instance, an autonomous vehicle must not only be endowed with an artificial morality which affords it an action-guiding recommendation in the case of a dilemma between a vehicle and a single pedestrian, but also any unavoidable collision scenario which could conceivably occur in the vehicle's environment: AV v. multiple pedestrians, occupied AV v. School Bus, occupied AV v. The Pope Mobile, etc. This, we will claim, demands something of a *generalizability condition* within the types of features and values we could hold to be morally salient; one which ensures that what we hold to be a morally relevant feature in one context must also remain relevant in most others endemic to an *Umwelt*.

[69] Or, under a particularly 'tight' interpretation of theory, human agents and sentient animals (singer, 1972).
[70] Nallur, 2020. This relates to the concept of 'domain robustness', where "Specific domains have their own moral desiderata, and a machine operating in a specific domain should be able to meet that domain's demands" (Bauer, 2020, 3). Ambiguity concerning the specificity of what can count as a single 'domain' however, has led some authors to advocate for the need for either a) general ethical principles which could be seen to be applicable across many domains (Bauer, 2020), or b) the conclusion that many proposals for ethical implementations in machine ethics fail the test of domain robustness (Nallur, 2020). Since we are concerned here only with the process of artificial moral uptake, rather than the ideal moral theory for use in artificial morality, we will leave these concerns to one side.

Put another way, the problem of artificial moral uptake forbids a strongly particularist or casuistic approach to the design of artificial morality. As Conitzer et al. maintain:

> When we try to classify a given action in a given moral dilemma as morally right or wrong…we can try to do so based on various *features* (or *attributes*) of the action. In a restricted domain, it may be relatively clear what the relevant features are…Even in these scenarios, identifying *all* the relevant features may not be easy…However, the primary goal of a *general* framework of moral decision-making is to identify abstract features that apply across domains, rather than to identify every nuanced feature that is potentially relevant to isolated scenarios.

To put a rather crude stamp on this idea then, we might say that the problem of artificial moral uptake is not one of pure casuistry, but rather pure *contextry*: the designer must identify the total set of morally relevant features and values that can be *generalized* across *every decision context* inherent to an *Umwelt,* and correspondingly, must identify those contextual features which a) are *never* seen to have moral importance[71], and b) whose moral importance is not sufficiently frequent to count as *generally* morally relevant. This idea has been addressed in the literature, often under the ominous moniker of the *moral frame problem*[72]. While it is true that the problem of artificial moral uptake shares some similarities with the frame problem as it is construed in classical AI, such an association fails to capture the true density of the problem. Indeed, Conitzer et al.'s treatment of this distinction between the general and particular moral relevance of a feature appears to be pragmatic in nature. This is to say that plausibly, there exists some threshold of moral uptake beyond which the identification of further features yields diminishing marginal returns, owing mainly to their rarity in the empirical reality of an *Umwelt*.

[71] We can see a parallel between what we have called *autonomy as provision* in chapter II (in which a human programmer anticipates every potential flux in an agent's environment and codes a solution to it) and the problem of artificial moral uptake, which aims at the same robustness on a moral level. Unsurprisingly then, the latter is often just as challenging as the former.

[72] Abney, 2012, 45: Beavers, 2011, 335. Abney's use of the moral frame problem focuses mainly on the problem of knowing what information is (ir)relevant to ethical decision-making in AMAs—thus capturing the selection of moral features—while Beavers is concerned with the scope of mainstream ethical theories in their application to artificial morality, a point we will address at multiple points in this section of the thesis, while it is a bit orthogonal here. Perhaps the tightest similarity to our vision of the problem of artificial moral uptake comes from Arkin (2009, 69), when he describes the challenge of artificial morality design as a question of what content needs to be represented to ensure the ethical application of lethality, and how to represent that content in the AMA's architecture.

Rather than debate this claim directly, it should be decidedly more interesting to search out other, acceptability-oriented reasons for the existence of such a threshold. To this end, the next section provides an illustration of how this problem is addressed by a designer, and how this activity is hampered by the notion of acceptability as institutional viability, with the help of another fictitious robot, the SoupSaint2020™.

3.1 *An Illustration of the Problem of Artificial Moral Uptake: SoupSaint2020™*

In keeping with our analysis thus far, it should first be useful to describe SoupSaint2020™ according to its HMAMA ontology, and to provide a bit of context:

> **SoupSaint2020™**: In the midst of a global pandemic, Bob, a designer at Baltimore Dynamics, is tasked with the design of a special artificial agent. In effect, mandatory social distancing measures have taken hold, and have further marginalized the homeless populations of major cities, barring their access to soup kitchens and shelters as the pandemic rages on. To combat this problem, Bob comes up with the design for SoupSaint2020™: a mobile soup-serving robot which ambles down the empty city streets, distributing hot chicken noodle soup directly to homeless individuals as it wanders across their path. Importantly, SoupSaint2020™ only operates *after* the government-mandated curfew, and thus operates on the assumption that any human agent it encounters who is not wearing a uniform *must* be a homeless individual. At first, Bob did not establish any particular order of priority amongst soup-claimants, operating instead under the standard norm of a 'first encountered, first served basis'. However, after troubling reports of disproportionality in soup distribution, and with certain homeless individuals failing to receive soup, Bob the designer must optimize SoupSaint2020™'s distribution, by designing a decision procedure which insures a morally responsive soup distribution across the city's homeless population.

To begin, we can clearly identify both the extensional and agentive restrictions of SoupSaint2020™: to deliver soup, and homeless individuals, respectively. This implies that the marks of significance in SoupSaint2020™'s *Umwelt* are limited to homeless individuals, and since we have presupposed that these individuals are the *only* human agents in the environment—thanks to the government curfew—we can avoid the complicated and contentious business of defining

which types of people, or which features 'count' as or indicate a homeless individual. Instead, we will focus on which features could be seen to inform a morally responsive distribution of soup amongst what we will call *soup claimants*, or those homeless individuals who could be the locus of moral responsiveness for SoupSaint2020™, and thus hold claims over its distribution[73]. Finally, since what concerns us here is the process of artificial moral uptake, and not the specific forms of moral uptake that standard moral theories require, we will initially proceed with the selection of features in a *a*theoretical fashion, looking instead at those features which likely figure in our considered moral judgements concerning SoupSaint2020™'s ideal behavior.

As we have maintained previously, the problem of artificial moral uptake demands the passage from a taxonomy of possible features in the AMA's environment, to a refinement towards what we might call *admissible* features: those which could be seen to have moral relevance, or to be salient to moral responsiveness. Importantly, because the moral responsiveness of an AMA is necessarily instantiated in its practical agency—or its decision-making capacity—we can view this process as one which investigates the morally relevant features of *option-context pairs*[74]: the decisional options or choices an AMA has within a specific context. Seen this way, any admissible feature must fall into one of three categories: (1) an *option* feature, wherein its possession by an option-context pair depends only on the option, and not on the context, (2) a *context* feature, wherein its possession by an option-context pair relies only on the context and not on an option, and logically, (3) a *relational* feature, wherein its possession by an option-context pair depends on *both* the option and the context.

We can sketch a quick example of these concepts by imagining a person who must decide which kind of coffee to order at a coffee shop[75]. Option features would be things like types of coffee beans and ways to make coffee (Robusta, Arabica, ristretto, allongé, americano etc.). These

[73] We will leave to one side the tempting *procedural* aspects of moral responsiveness, for instance, the idea of ensuring a fair distribution amongst soup claimants through the establishment of procedural rules like 'one soup per person per day'. These will crop up in our more robust discussions of artificial morality throughout this section of the thesis, but they distract us from our current problem, and thus we will assume that these types of decision procedures are already in place.
[74] Dietrich & List, 2017.
[75] Dietrich & List, from which these distinctions are borrowed, sketch a similar example of choosing what to order at a restaurant (2017).

features are option features because they would not vary across contexts: an allongé is an allongé regardless of the coffee shop one frequents (the context). Context features would be things like the price of coffee, the number of menu options, 2 for 1 deals or special sales. These do not relate to the specific types of coffee from which one may choose, but instead, are specific to a given coffee shop (the context). Finally, relational features would likely include relative aspects such as cheapest, sweetest, strongest, highest caffeine content, or least calorific. All of these features depend both on the options (types of coffee) and the context (which kinds are available at the shop), since the sweetest coffee at one shop may be the most bitter at another, and so on. Of course, many human agents perform 'coffee casuistry' on a daily basis: through the identification of admissible features which together combine to expound the agent's taste in coffee— i.e. cheapest, most caffeine content, arabica beans, and an americano—when ordering at their favorite café. Indeed, it is even plausible that some unfortunate human agents occasionally face 'coffee dilemmas', say, when they are forced to buy coffee from an automated distributor which offers only two options: cheapest and ristretto, or decaf and americano. Importantly however, it appears difficult to establish a *generalized account* of the admissible features of coffee selection—or those features which could hold regardless of the café or automated distributor from which we choose, or the types of coffee available. Nevertheless, a response to the problem of artificial moral uptake (and thus the design of artificial morality) aims to accomplish precisely this in an AMA's *Umwelt*[76].

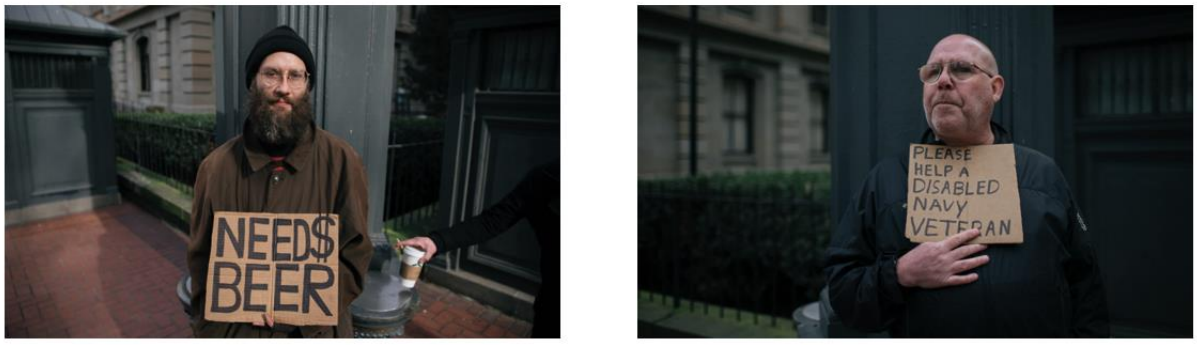
With these terms more firmly established then, we can return to our example of the SoupSaint2020™. What types of admissible features could be seen to apply to an agent tasked with the distribution of soup amongst soup-claimants? Immediately, we should recognize that the SoupSaint2020™ must make decisions under *scarcity*: it does not have an unlimited supply of soup to offer to the (plausibly) *unlimited* demands of soup claimants[77]. In this sense then, it may

[76] We must recall however that an AMA, as opposed to a human agent, only operates within a limited environment or *Umwelt*, and is for this reason not a universal practical agent. In this sense, the contexts within which the AMA will be expected to act (which are themselves further truncated by the AMA's extensional restrictions) form a comparatively smaller set than those of human agents. Accordingly, it is more reasonable to assume that most of the ethically salient features we should wish to select could be generalized across all of the decision-contexts in the agent's *Umwelt*.

[77] In economic terms, this is loosely tantamount to *supply-induced* scarcity, where the supply is significantly lower than the demand for a particular good.

be the case that not everyone's soup claim can be satisfied by the SoupSaint2020™. This, in turn, has two consequences for the SoupSaint2020™'s practical agency: first, that it may *necessarily fail* to respond to some soup-claims, and thus cause some degree of *harm* to these claim holders; and second, that it may need to establish a *ranking*, order of priority, or some type of *trade-off* between the soup-claimants, so as to decide which of the soup claims will be satisfied, and which will not. Thus, in our quest to understand which features of SoupSaint2020™'s environment ought to matter morally, we should bear in mind that these resultant admissible features will serve a *discriminatory* role in the AMA's decision-making, causing the machine to decide to distribute soup to some individuals rather than others by shaping its response to moral value.

To jumpstart our selection of admissible features, we can imagine a scenario wherein SoupSaint2020™ has only one unit of soup remaining, and must decide which of two claimants ought to receive the soup. Let us suppose that the individuals depicted in the figure below represent these two claimants:



claimant 1 *claimant 2*

Fig. 1 - *Two Soup Claimants*

Intuitively, many of us would hold that the last remaining unit of soup should go to claimant (2). We might even say that common sense morality would support the choice of claimant (2) over claimant (1). Yet while many human decision-makers may choose in this way, the designer of an AMA, in his response to such a problem of artificial moral uptake, must be able to provide a *generalized* account of the morally relevant features which would prompt the choice to privilege (2) over (1). In this way, when encountering a similar soup dilemma, SoupSaint2020™ will

respond in a similar way, displaying a *consistent* response to the moral value of its environment. Starting with only option features, we could identify the following characteristics across these two claimants: age, height, weight, gender and demeanor. If we are willing to briefly divorce from the technical limits of machine perception—as is often done in trolley cases[78]—we might also include features such as socio-economic status, medical history, criminal background, family ties, educational background, political or religious ties or other such popular markers of moral interest in moral philosophy. Contextual features, in turn, would reduce to such aspects as the number of claimants, the latest date and time at which a particular claimant received soup, weather conditions, or distance between claimants. Finally, we could find myriad relational features which could ostensibly be useful in deciding between claimant (2) and (1): more physical comparisons such as the oldest, the tallest, the healthiest, the hairiest, or the claimant with the warmest clothes; while we could also find a much longer list of open-textured and somewhat dubious comparisons: the hungriest, the nicest, the friendliest, the most vulnerable, and of course, the most deserving.

It is obviously tempting to lean heavily on relational properties like vulnerability or desert, and to ignore the trickier aspects of deciding which option features ought to matter morally, or could be seen to track particular moral values. But beyond the apparent circularity inherent to the use of a vague relational feature like 'most deserving'—which would indeed require the assessment of desert via the perception of a number of relevant option and context properties—at a generalizable level, this choice runs abreast of the frame problem in classical AI due to its reliance on *ceteris paribus* reasoning: other things being equal, claimant (2) is more deserving. Exactly *which* things are being held equal, reduce to precisely the context and option features which we sought to avoid[79].

[78] It is quite clear that most autonomous vehicles—barring impressive improvements in vehicle-to-vehicle or vehicle-to-device communication—would not be able to reliably intuit an individual's socio-economic status or his criminal background, despite these being the types of admissible properties which the MIT Moral Machine Experiment sought to exploit in their study (Bonnefon et al., 2016).

[79] Daniel Dennett, in thinking about the frame problem writes: "The beauty of the ceteris paribus clause in a bit of reasoning is that one does not have to say exactly what it means…If one had to answer such a question, invoking the ceteris paribus clause would be pointless, for it is precisely in order to evade that task that one uses it. If one could answer that question, one wouldn't need to invoke the clause in the first place. One way of viewing the frame problem then, is as the attempt to get a computer to avail itself of this distinctively human style of mental operation" (1998a, 198).

Thus, the identification and subsequent ranking of option and context features in the response to artificial moral uptake appears unavoidable, either as a means to the substantive definition of relational features like 'most deserving', or as a comparatively stand-alone construction of what ought to matter morally. In this sense, we might argue that the option features which make claimant (2) more appealing relate to three facts: the age of the claimant, the disability of the claimant, and perhaps the fact that he is a veteran. Using these facts, we can construct a number of plausible explanations for choosing claimant (2) over claimant (1)—all of which supporting some type of relational feature of SoupSaint2020™'s decision context. For instance, we might say that the age and disability of claimant (2) render him more *vulnerable* than claimant (1), and for this reason he should receive the soup. We might also say that claimant (2) has made a greater contribution to society (in virtue of his military service) than claimant (1), and thus that he should receive the soup. In this sense, we might be tempted to devise a 'disabled veteran trumps all' type of response to the problem of artificial moral uptake, which would likely track the sentiment of common sense morality, and perhaps some loose forms of Rawlsian theory, in so far as these individuals are likely the worst-off claimants in most decision contexts. What happens, however, when the SoupSaint2020™ is confronted with further soup claimants, as depicted in figure 2?

claimant 2 *claimant 3* *claimant 4*

Fig. 2 - *further soup claimants*

The choice between claimants (2), (3), and (4) is not aided in any meaningful way by the three option features which were seen to contribute to our original choice of claimant (2) over claimant (1). In this further case, all three claimants are veterans of advanced age, and all three are

disabled. In other words, if the soup claims of these individuals were grounded solely in these option features, it would not be clear *which* claimant ought to receive the soup. We would then be faced with two dubious solutions: first, we could further specify the relational feature of vulnerability—in this case bringing in the uncomfortable consideration of *how* disabled an individual claimant is compared to others. This would likely lead us to privilege the claim of either claimant (3) or (4), since (4) has no legs, but ostensibly has a more mobile wheelchair than (3). Or, we might hold that the extensional definition of the relational feature of vulnerability has reached its maximal point, and thus may opt for the randomization between claimants, in which case SoupSaint2020™ would select one claimant indiscriminately. We may hold that this is plausible given the unlikelihood of such a comparison occurring often in the SoupSaint2020™'s *Umwelt*. Importantly, we can see that the potential for equal claims in an AMA's *Umwelt* generates strong moral tension: we must either look for a further decisive fact (or feature) which provides a reason to privilege one claim over another—making this claim *stronger* than the others—or, we must concede that all claimants demand an equal degree of moral responsiveness, and thus randomize between these equal claims[80]. It is likely at precisely this point that the distinction between *general* morally relevant features, and features which are relevant to only *particular* contexts, becomes most poignant. The threshold, in this sense, is exemplified by an informational barrier beyond which it becomes appropriate to *randomize* a machine's moral responsiveness across claimants.

In this sense, we can clearly see that the choice of admissible features in an AMA's environment may lead us to make uncomfortable moral trade-offs, or at least, may tempt us to extensionally define otherwise acceptable norms and principles to an uncomfortable degree. To this, we should add two additional problems prevalent in the literature: the choice of which admissible features are *morally acceptable* regardless of their descriptive utility in an *Umwelt*[81], and the moral concern for the value trade-offs which the selection of certain admissible features may cause in the agent's *Umwelt.* As we maintained in the previous section, these concerns are perhaps given their most robust treatment in the literature on autonomous vehicle ethics, the paradigmatic example revolving around either a) the characteristics chosen by the Moral Machine

[80] Leben, 2017: Keeling, 2018.
[81] Evans et al., 2020: Jacques, 2019.

Experiment, or b) Noah Goodall's thought experiment framing the choice of selecting 'helmet-wearing' as an admissible option feature of an autonomous vehicle's response to artificial moral uptake[82].

Expanding now on the latter example, according to Goodall, the choice to discriminate between those cyclists who are wearing helmets and those who are not leads to a value trade-off, depending on how this feature contributes to the claim of an individual: if helmet-wearing is seen to contribute to a *decrease* in vulnerability, then autonomous vehicles may penalize or 'target' helmet-wearing, law-abiding citizens, by choosing to sacrifice these agents over non-helmet-wearing cyclists. This may generate an incentive amongst cyclists to refrain from wearing helmets, lest they be targeted by the autonomous vehicle in the case of an unavoidable accident. Undoubtedly, this situation is not optimal, as it would generate a higher risk of human harm in the traffic community at large[83]. Thus, it would seem that while helmet-wearing could be seen as an *admissible* option feature of an AV's *Umwelt*—i.e., in so far as the presence of this feature affects welfare—it may not be an *acceptable* feature, in light of the perverse incentives it may generate in the traffic environment. Let us call this the *perverse incentive problem*.

The notion of a perverse incentive, in turn, has been addressed in the autonomous vehicles literature, in an insightful article by Wolf Loh and Catrin Misselhorn[84]. In detail, the authors distinguish between two types of incentives which might be cause for moral alarm: what they call *adverse incentives*, "…the undesired side effects of actions, technologies, or social policies, which are directly incentivized by the relevant action, technology or social policy"[85]; and *perverse incentives*, which "…invite behavior which directly negates the primary goal of the said action,

[82] "If the collision is severe and injury is likely, the automated vehicle would choose to collide with the vehicle with the higher safety rating, or choose to collide with a helmeted motorcyclist instead of a helmet-less rider. Many would consider this unfair not only because of discrimination but also because those who paid for safety are targeted while those who did not are spared." (Goodall, 2014, 8). This same experiment is also discussed in (Lin, 2016).
[83] Of course, the inverse is likely just as undesirable: in choosing to target *non-helmet* wearing cyclists over those that uphold helmet laws, the autonomous vehicle could be seen to 'punish' those road users who do not follow the letter of the law, displaying a bizarre form of technological paternalism.
[84] Loh & Misselhorn, 2019.
[85] Ibid., p. 576.

technology or social policy"[86]. Since the authors view public safety as the primary goal of autonomous vehicle implementation, they view the helmet case as an example of a perverse incentive[87]. In this sense then, the perverse incentive problem points to a troubling tension between, on one hand, the types of admissible features that would track our considered moral judgements, or track the specific structure of a given moral theory; and on the other, the telos of the technology itself. Put another way, it would seem that a 'tight' or exhaustive interpretation of some popular morally relevant features—such as welfare or vulnerability—may undermine an artificial moral agent's moral responsiveness; cancelling-out, so to speak, many of the morally salient benefits (such as a reduction in harm) that the implementation of artificial morality is meant to confer on the behavior of the agent. Thus, the perverse incentive problem provides us with our first acceptability-oriented reason to place a threshold on the moral uptake of an artificial moral agent: so as to avoid or minimize any ethically salient blowback from the publicity of certain features may cause[88].

In an entirely different vein, there is a broader sense for which certain admissible features may prove to be unacceptable, one which aligns with what we have called ART principles[89], or more broadly, ethical design concerns. We can grasp this concern if we ask which features provide acceptable *decisive* reasons for the response (or non-response) to an individual's claim. To make things simple, we will call this the *further feature problem*. In figure 2, we saw that claimants (2), (3), and (4) were seen to hold equal claims over SoupSaint2020™'s last remaining unit of soup, since all three were disabled army veterans of advanced age. If we were hostile to the idea of

[86] Ibid., p. 576.
[87] "If (a) the crash algorithms of fully autonomous vehicles try to minimize harm in an accident, (b) the motorcyclists know about this fact, and (c) fully autonomous vehicles become one of the prevailing traffic participants, it may become rational for motorcyclists to drive without a helmet. In this case, the implementation of these kinds of crash algorithms incentivizes a rational change in behavior, which in turn may lead to more serious self-inflicted accidents" (Loh & Misselhorn, 2019, 579).
[88] In effect, the perverse incentive problem, and to a certain extent, the arguments of Low and Misselhorn, presuppose a certain publicity condition, wherein individuals (other than the passenger) are aware of the details of an autonomous vehicle's artificial morality, and can therefore adjust their behavior accordingly. In the case of autonomous vehicles at least, it must be admitted that the plausibility of this publicity condition is somewhat questionable given the relatively tight-lipped behavior of original equipment manufacturers in this regard. However, a firm adherence to the principle of transparency in the design of AMAs generally would likely recommend that such information be publicly available, and thus it is nevertheless plausible that this condition could be met.
[89] Dignum, 2019.

randomization, then we were forced to find a further feature amongst these claimants which breaks the dead lock between their three equal claims. This drove us to further specify the relational property 'vulnerability', ostensibly by evaluating the *number of legs* of each claimant, contentiously choosing claimant (3) or (4). Imagine now that, in line with the principle of accountability, some further human agent pressed SoupSaint2020™'s 'why did you do that?' button. It would necessarily reply something to the tune of: 'because claimant (4) had less legs than claimant (2) or (3)'. This, to the outside observer at least, is hardly an acceptable reason. Thus, this is one dramatic example of the further feature problem: while certain features may have decisional utility in an agent's *Umwelt*, they may not provide acceptable grounds for depriving or awarding responsiveness to human agents. They may not track, in other words, 'background constraints' such as human rights and dignity, or broader data protection laws or privacy concerns. In a broader sense, other aspects such as criminal background, health history or socio-economic status, while being useful distinctions amongst claimants, may nevertheless fail to be acceptable distinctions for similar reasons. Put simply, even if these features may figure in a human agent's appraisal of the moral status of an act, their use in *artificial morality* often thwarts ethical design concerns—and thus seem to represent facts which it is unacceptable for a machine to know about an individual human being.

Importantly, the further feature problem affects the problem of artificial moral uptake in two interrelated ways. First and most generally, it can mandate a very liberal application of what we might call a *threshold of moral blindness*: that degree of informational specificity for which the inclusion of any further fact or characteristic is unethical (in virtue of ethical design concerns), and thus the point at which an AMA ought to *randomize* across claimants, or equally strong moral claims. This threshold of moral blindness differs from the threshold provided by the problem of perverse incentives, however, in so far as the latter addresses the *negative externalities* of a particular morally relevant feature on the empirical reality of an *Umwelt*, while the former addresses the *incompatibility* of a morally relevant feature with pre-established design norms, such as those expounded by ethical design concerns[90]. In this sense, the further feature problem presents

[90] These two thresholds differ also in their reliance on a publicity condition. In the case of the threshold of moral blindness, it is not necessary that the public itself be aware of the details of an AMA's artificial morality, only that some organization or institution is aware of its violation of ethical design concerns.

a designer with a different reason to install an informational threshold in an AMA's artificial morality: adherence to ethical design concerns.

In the case of the SoupSaint2020™ then, we rather intuitively maintained that the further fact of the 'number of legs' of a claimant seemed to exceed the threshold of moral blindness, and in this sense, it seemed appropriate to randomize across the three claimants in figure two, all of which were disabled army veterans of an advanced age. Depending on the application domain and the type of computational decision procedure, the appropriate level of this threshold will vary widely. Indeed, in the case of bottom-up approaches, concepts such as privacy by design[91], or computational constraints such as k-anonymity[92], can be seen as instantiations of a threshold of moral blindness, insofar as they shield an agent program from useful but unethical information which might otherwise inform its decisions. Importantly, the stringency of the threshold of moral blindness has an inverse relationship with decisional efficiency. This is to say that the less a machine 'knows' about its decision context, the more uninformed or potentially inaccurate[93] its decisions will be, from a pure technical standpoint. Somewhat paradoxically then, it seems to follow that a highly ethical machine from an internal perspective—one whose artificial morality admits many morally relevant features, and thus rarely randomizes across individuals or actions—may be highly *unethical* from the external standpoint of ethical design concerns, and vice versa.

Secondly, the further feature problem can be seen to exert a *downward pressure* on both the selection of admissible features in an AMA's *Umwelt, and* the shape of its computational decision procedure; seeking to minimize, if not erase entirely, a number of ostensibly viable criterions of rightness, procedures, and features which could serve in the decisions of distributive AMAs. The most obvious example of this line of thinking is found in the German Federal Ministry

[91] Langheinrich, 2001.

[92] *K*-Anonymity is a tactic used to "…redact information from individual records so that no set of characteristics matches just a single data record. Individual characteristics are divided into 'sensitive' and 'insensitive' attributes…[where] the goal of *k*-anonymity is to make it hard to link insensitive attributes to sensitive attributes" (Kearns & Roth, 2019, 27-28). See also (Sweeney, 2002).

[93] "The first major consequence [of ethical constraints] is that we will now have algorithms that are guaranteed to have the particular ethical behaviors we asked for. But the second major consequence is that these guarantees will come at a cost—namely, a cost in the accuracy of the models we learn. If the most accurate model for predicting loan repayment is racially biased, then, by definition, eradicating that bias results in a less accurate model" (Kearns & Roth, 2019, 18).

of Transport and Digital Infrastructure's Ethics Commission on Automated and Connected Driving[94], which stipulates:

> In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties[95].

This would seem to suggest that the *only* admissible feature in an autonomous vehicle's artificial moral uptake is the context feature 'number of claimants', or the relational feature 'least personal injuries'. This yields an autonomous vehicle that aims to assess individual claims in collision scenarios *only* by evaluating the number of personal injuries which are liable to result from the pursuit of a given action[96]. Further, the German commission's recommendation removes the possibility of a distinct number of criterions of rightness and theoretical decision procedures from use in autonomous vehicles, barring for instance, the possibility that one claimant can be sacrificed so as to ensure the survival of another, or plausibly, that by-standers such as surrounding pedestrians must not be harmed as a result of the vehicle's decision. Instead, the proper criterion of rightness and theoretical decision procedure appears to be limited to the relational property of 'causing the least amount of casualties'. From one side, this seems to point to a highly deontological account of artificial morality, or one that only holds agent-centered constraints such as 'do no harm' as normatively relevant factors. From the other side, and somewhat paradoxically, this particular set of constraints seems to forbid the possibility of a 'general reduction in personal injuries'. In other words, the criterion of rightness and the decision procedure recommended by the commission seem to be mutually incompatible without further specification. Generally however, in ways that differ from some of the forms of artificial morality we have thus far

[94] Luetge, 2017: BMVI, 2017.
[95] BMVI, 2017, 7.
[96] To complicate things further, the utilitarian calculus that is presupposed by this type of assessment is thwarted by the commission's stipulation that it is forbidden to 'offset victims against one another', which we might loosely understand as the prohibition of using some human agents as a means to the survival of others, with some degree of Kantian inspiration. It would be forbidden, in this sense, to respond in any way to unavoidable collision scenarios, since these necessarily offset individual claimants against one another. Of course, given the commissions avowal to avoid (at all costs) these scenarios in the first place, it should not be too surprising that its recommendations are so stringent.

explored, this particular account provides an incredibly minimal response to the problem of artificial moral uptake, albeit one that evades the further feature problem, and maximally aligns with ethical design concerns.

Interestingly however, the Commission's prohibition of any personal features in the process of artificial moral uptake likely tacitly assumes that option features such as age and gender are intimately and *exclusively* linked to a given individual's identity, and thus, that the use of these features in AMA decision-making opens the system up to concerns of discrimination, or thwarts background constraints such as human rights, equality and dignity. In one sense, this is true, but in another, important sense, it may not be. To understand why, we must posit a distinction of sorts between two types of facts which may come to be admissible features in an AMA's *Umwelt,* what we will call *constitutive facts* and *scalar facts.*

In this section, we have seen that intuitively desirable relational features, such as 'most deserving' or 'most vulnerable' require a substantive specification of the option and context features which together provide a generalizable account of what allows the machine to detect 'vulnerability' or 'desert' in any context within its *Umwelt*. In simpler terms: we cannot avoid substantively defining relational features in general terms, via the option and context features which support them. As the complexity of the AMA's environment increases, or as it comes into contact with claimants who seem to equally satisfy our posited option and context features, we will need to search out *further features* which usefully disrupt this equality amongst claims. In SoupSaint2020™'s case, this further feature was the number of legs of a given claimant, transforming the relational feature of 'vulnerability' into the relational feature 'least amount of legs' in contexts where all claimants are disabled army veterans of an advanced age. What kind of a fact is this?

Clearly, the number of legs of a given claimant counts as an *option feature*, since a given claimant will have the same number of legs regardless of the context in which he encounters SoupSaint2020™. In one sense then, we could say that the number of legs of a given claimant is a *constitutive fact*: it is a *constitutive* element of the concept of vulnerability in SoupSaint2020™'s *Umwelt*. In this sense, we may also view *age* as a constitutive element of the concept of

vulnerability, if we can agree that advanced age renders an individual more vulnerable to the risks or dangers of life on the streets. Thus, constitutive facts serve to bolster desirable relational properties like vulnerability or merit[97]. Accordingly, when a designer sets out to construct a response to the problem of artificial moral uptake, he will seek out those constitutive facts which support the relational features he finds desirable: baldness is constitutive of 'the hairiest', wearing-a-jacket is constitutive of 'the warmest clothes', smiling is constitutive of 'the nicest', and so on.

When employing a particular *moral theory* in the design of artificial morality, the designers approach occurs in three steps: identify the normatively relevant factors of the theory (i.e. general utility), identify the morally relevant features which could be seen to flow from these factors in a particular *Umwelt* (i.e. welfare, risk of harm or death), and finally, identify the marks of significance (option, context or relational facts) which are *constitutive* of these morally relevant features (i.e. the safety rating of a particular individual's vehicle, whether or not the cyclist is wearing a helmet, etc.) In this sense, moral tension arises when the designer chooses a constitutive fact which is a) insufficiently constitutive of a desirable relational property, b) liable to generate perverse incentives in real-time interaction, or c) itself contentious or disrespectful of ethical design concerns and background constraints, such as the 'number of legs' of a given claimant.

In this way, the German Commission's prohibition of personal features such as age and gender is clearly driven by both (a) and (c): they deny that age and gender could be constitutive of their desired relational property, 'least amount of personal injuries' in AV collision scenarios; and they likely find these features to be contentious, disrespectful or unjustifiably discriminatory. Instead, they may view features such as age and gender as what we could call *scalar facts*: indications of how tightly an AMA's actions track the 'wisdom of the crowd', normative expectations, or individual and societal acceptability. Scalar facts, in other words, pertain to the realm of descriptive ethics: facts which point to the moral attitudes, beliefs and characteristics of groups or individuals. In this sense, the MIT Moral Machine Experiment can be seen as one, enormously expansive attempt to discover the scalar facts which surround the ethics of autonomous vehicle decision-making in unavoidable collision scenarios. Accordingly, it is a scalar

[97] As we shall see in chapter VII, they also constitute or operationalize values, norms and moral principles, themselves indicating certain desirable relational properties.

fact that France prefers to spare the young over the old, and females over males[98], while Germany comparatively prefers the elderly and males. Scalar facts then, seem to almost necessarily point to the inherent biases and preferences of sampled populations, and in this sense, it is no small wonder why the German Commission wished to avoid incorporating these into formal policy.

Nevertheless, in the case of autonomous vehicles at least, it should be somewhat clear how both age and gender could be considered as *both* scalar and constitutive facts. If individuals of advanced age are more likely to be injured as a result of a collision with an autonomous vehicle[99], or more likely to be injured in severe ways, then age is a constitutive fact of the German commission's preferred relational feature, 'least personal injuries'. Similarly, if women are disproportionately injured in collisions[100], then gender too could be a constitutive fact. Thus, the prohibition of any and all 'personal features' in a designer's response to the problem of artificial moral uptake may come with risks: most saliently, it may prohibit the useful operationalization (or constitution) of a desirable point of comparison across claimants, and in so doing, fail to provide an appropriate amount of moral responsiveness in AMA behavior.

To summarize, our investigation of the problem of artificial moral uptake, and the SoupSaint2020™, has exposed a number of ways in which the identification of what matters morally proves exceedingly difficult in most AMA *Umwelts*. From the identification of the option, context and relational features which exist in the AMAs environment, there may be moral or epistemic uncertainty as to which of these could constitute admissible features in the AMA's artificial morality. To this, we must add the distinction between general and particular features, and how this resulting informational threshold is modified by the perverse incentive problem, and the further feature problem. While the imposition of the former staves off ethically dubious outcomes in practice, the latter places serious limits on what a machine should know in light of

[98] Awad et al., 2019. Indeed, out of 117 countries, France ranked 1st and 2nd in these areas, respectively. A cooperative perusal of the experiment's results can be found here: http://moralmachineresults.scalablecoop.org/

[99] Kim et al., 2008.

[100] This idea has garnered empirical support, as well as the telling name of the *danger divide* between male and female crash victims. Curtly, empirical studies have shown that in virtue of the fact that many crash dummies are modelled to reflect the male form, females are 47% more likely to suffer severe injuries in car crashes (Bose et al., 2011).

ethical design concerns, both of which have serious impacts on the overall efficiency of an AMA's artificial morality. Further still, more exhaustive institutional recommendations may exert a downward pressure on the design of artificial morality: removing minimally the possibility of certain morally relevant features and constitutive facts, or maximally, a host of viable decision procedures and criterions of rightness for a given type of AMA, mainly by considering only certain relational properties as admissible forms of discrimination.

Generally then, recognition of the types of constraints that acceptability as institutional viability can pose, may prompt the suppression of many, (especially option) features which could be seen to matter in a given environment, generating something of a technological catch-22: too many features, and the machine is liable to operate on unacceptable motives in its moral responsiveness, but too few, and the machine may fail to properly apprehend the moral value of its environment, and thus generate a relatively unresponsive artificial moral agent. Finally, this process is further complicated by the determination of constitutive and scalar facts, where many basic, useful features can conceivably pertain to both. Here again, the exclusion of a constitutive fact (due to its mis-categorization as scalar) will adversely affect the clarity and precision of the machine's vision of the moral value in its environment; blurring or rendering it blind to important features which would likely contribute to optimal ethical decision-making. However, the inclusion of scalar facts may be the harbinger of machine bias, and generalized discrimination, and thus cause the machine to disproportionately or unfairly respond to the claims of certain human agents over others.

In this sense, it is important to understand, when investigating which option and context features underpin the desirable relational features of an AMA's environment, *how* these features bolster comparisons across claimants. It is possible, for instance, to use a feature such as age in a constitutive sense, without incorporating the bias its scalar use may engender. It is also possible that increased exposure to these machines, and increased transparency and justification concerning how they are programmed, might round off some of the sharper conclusions we have drawn here. Nevertheless, we are now quite aware of the complications inherent to *any* response to the problem of artificial moral uptake, even in its simplest form.

4. *Conclusion*

We began our exploration of the concept of acceptability with the simple claim that the humanistic intensions behind many (current) artificial moral agents appear to require us to take acceptability seriously in AMA design. From there, the brunt of our analysis was spent deciphering three distinct notions of acceptability, understood either as the expression of moral preference, the adoptability of an AMA, or an AMA's institutional viability. Each of these notions seemed to chip away at both the authority and structure of artificial morality understood as an exercise in moral philosophy, where the computational implementation of a moral theory is met with specific and non-negligible constraints. In the case of acceptability understood as the expression of moral preference, we understood that empirical research may provide both a) the geographical limits of the acceptability of a given type of artificial morality, and more importantly b) may expose the shape of local common sense morality, providing a view of how an acceptable AMA ought to reason in specific contexts. From the notion of acceptability as adoptability, we gleaned the importance of user satisfaction in particularly *surrogate* forms of AMA, and how a failure to account for user preferences, even if this runs against the more maximalist aspirations we might have for artificial moral agents, may prevent an AMA from achieving the ubiquity necessary to act 'better than' a human agent, and thus bring about the 'best' or 'correct' results. Finally, from acceptability as institutional viability, we understood how the imposition of a high threshold of informational blindness and a failure to differentiate between the scalar and constitutive facts of an AMA's *Umwelt* together forbid a vast array of potential morally relevant features, decision procedures, criterions of rightness or normatively relevant factors. In extreme cases, this may severely impair the moral responsiveness of an artificial moral agent, or efface the decisional utility of many types of moral theory all together. In simpler terms, it appears that the position of taking acceptability seriously affords very little space for pure moral reasoning in AMAs, and even less space for moral theory itself. To this end, the following chapter attempts to address the vacant space left over by both acceptability and technical constraints, and to devise a kind of artificial morality which can nevertheless yield morally responsive artificial moral agents.

CHAPTER

6

Artificial Morality & The Ethical Valence Theory

At the outset of part II of this thesis, we introduced what we called the 'diamond question' of machine ethics. This question asked which criterion of rightness, theoretical decision procedure, and computational decision procedure together provide an agent program that successfully moves an ethically optimized machine that acts like a moral agent from perception to action in a specific *Umwelt*. In many respects, the arguments of chapter IV and V painted this challenge as a sort of rivalry between what we have called the *maximalists* and the *minimalists*. Despite their common goal of building explicit ethical agents, these two factions were seen to disagree on one very fundamental issue: the *source* of moral content which was meant to inform the agent's behavior. The maximalist expounded the stringent use of (a particular) moral theory, and in so doing, advocated for the maximal application of normative ethics, even at the cost of thwarting the agent's *practical push.* In this sense, the maximalist view was championed by the Philanthropotron3000™, a purposeless agent upon which the demands of morality know no bounds. On the other side, the minimalist advocated for the use of empirical data or the methods of descriptive ethics as the ideal source to inform robotic behavior, data which may only minimally track the rough contours of

some recognizable ethical paradigm. The minimalist, in turn, might then be championed by what we could call the Parochiatron3000™, whose extreme mobilization of local common-sense morality risked the perpetuation of bias, instinct, unreflective preferences or worse[1].

It would then seem that any response to the diamond question forces us to choose between these two conceptual strawmen, despite their being quite unpalatable ideals[2]. Further still, neither model seemed to escape the hurdles posed by technical and acceptability constraints unscathed. The maximalist was forced to contend with the 'pushless' quality of artificial moral agents, a fact which, when taken seriously, lead to the idea that these agents could not figure among the constituents of any moral theory, and thus that an important degree of theoretical interpretation was required to afford them any action-guiding recommendations at all. In this way, much of the initial value, and ethical security which flowed from the strict implementation of these theories seemed to be threatened. The minimalist, on the other hand, was forced to accomplish his own form of interpretation when attempting to force his empirical data into the mold of a particular moral theory; or, by delegating this interpretation to the machine itself, generated significant problems of coherence and opacity.

Finally, both factions suffered from the challenges inherent to the problem of artificial moral uptake, since the extensional definition of a principle, or the instantiation of a morally relevant feature, both ran up against the problem of perverse incentives, and the further feature problem. In this sense, a threshold of moral blindness might reasonably prevent the maximalist

[1] A particularly lyrical example of this concern is expressed by Derek Leben: "Think of all the terrible practices and institutions in human history: slavery, genocide, caste systems, public torture, and so on. If robots had existed at the time, we would have wanted them to refuse to participate in these practices, and perhaps even prevent humans from continuing them…It is not only possible but likely that future generations will look back in horror at many of the practices we are currently engaged in (factory farms, fossil fuel burning, prison systems, massive inequality, etc.). Robots that have no moral principles, or the wrong moral principles, will only make these injustices more efficient" (Leben, 2018, 147). On a more individualistic bent, J Storrs Hall argues that the implementation of more utilitarian forms of artificial morality may help flatten the curve of partiality—or what he calls the 'sombrero of moral concern'—in common-sense morality (Hall, 2009, 303-308).

[2] This assumes of course, that either route is still worth pursuing, an idea of which some machine ethicists remain doubtful: "Any illusion that fundamental unit of moral value can be divorced from its place in the world, inward to modular mechanisms or outward to rationalist principles merely distracts from the hard toil of becoming a moral person through action, and it is at the interface of entity and environment that this work takes place" (White, 2014, 372).

from detecting those features which are required for the optimal performance of a theoretical decision procedure; while the inability to mobilize many ethically charged features in the empirical research of the minimalist may hamper his ability to illicit the intricacies of the wisdom of the crowd. The figure below illustrates these problems:

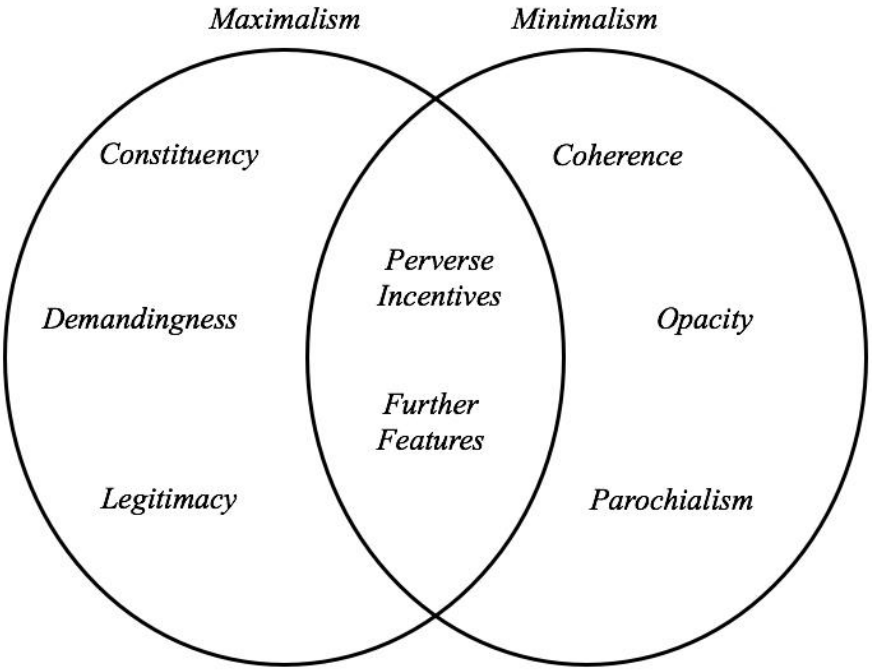


Fig. 1— *the problems of minimalism and maximalism*

Save for the problems of *legitimacy* and *parochialism*, the themes listed in the figure above should seem familiar given the arguments of part II. This is so since these last two problems do not relate to technical or acceptability constraints, but rather, relate more to the fundamental purpose of artificial morality itself. To see why this is the case, let us briefly introduce these concepts. The problem of parochialism, as our brief introduction has alluded to, relates to the minimalist's necessary commitment to the normative preferences of whichever population provides his source of moral content. This might be an individual end-user in the case of customizable ethics settings or 'ethics bots'[3], or millions of serious game participants as in the case of the Moral Machine Experiment. The common claim to make faced with this problem is an

[3] Etzioni & Etzioni, 2017.

appeal to the biases or injustices that could naturally result from such a heavy dependence on this source of moral content[4]. This is, for all intents and purposes, the problem of parochialism. However, much of the force of this claim depends on the minimalist's misalignment with ethical design concerns, and thus often presupposes that certain unacceptable further features will figure in the world knowledge of the artificial moral agent, and accordingly be reflected in its moral decision-making.

Let us, for the moment, remove this barrier by maintaining that minimalist approaches to artificial morality can overcome the problem of further features while still yielding workable models of artificial morality, and thus align themselves with ethical design concerns. What we are left with is a source of moral content which reflects a somewhat idealized shape of common-sense morality particular to a given *Umwelt*[5]. In this sense, it is fair to assume that this approach will capture a very wide range of the sample population's considered moral judgements. Or, that the resulting action-guiding recommendations will be sensitive to a wide range of the normative expectations that many individuals hold as to a given AMA's behavior: a general concern for utility, a special concern for users interacting closely with the machine, certain injunctions concerning harming or deceiving human beings, etc. Robots, when programmed this way, will likely behave in ways that are conducive to the expectations of human agents, or in any case, the mode of their moral responsiveness will appear neither abhorrent nor heroic. Let us call such an achievement the *minimalist's ideal*.

Consider now what we will call the problem of *legitimacy* for the maximalist. This problem relates to the inherent *arbitrariness* of the choice of moral theory for implementation, and it is a pervasive problem in machine ethics. We can grasp the force of this problem if we ask what reasons we would have to implement a Kantian form of artificial morality, rather than a Smithian or a Rawlsian model into a specific AMA? In effect, it would seem that there are only three routes by which the maximalist can answer this charge: first, that the designer somehow increase his

[4] Tolmeijer et al., 2020: Bonnefon et al., 2016: Awad et al., 2018: Leben, 2018: Jacques, 2019: Lin, 2016: Bauer, 2020: Nallur, 2020: Conitzer et al., 2017: Dressel & Farid, 2018: Etzioni & Etzioni, 2017: Goodall, 2019: Gabriel, 2020: Jobin et al., 2019.

[5] Thereby avoiding another critique of minimalist approaches related to their inapplicability to contexts larger (or different) than those upon which their data was collected (Winfield, 2019).

epistemic certainty that a given moral theory is correct *sub specie aeternitatis*, thereby vindicating his theory in the outright[6]; second, by maintaining that the given theory is relatively correct given that the minimalist's data on moral preferences tends towards this ideal; or finally, by holding that the recommendations of a given moral theory align with the practices endemic to its *Umwelt*, as might be the case in warfare or healthcare[7]. Notice then that barring a stout espousal of moral realism, each of the maximalist's options require some justificatory force from the normative expectations of human agents[8]. Or, that the choice of moral theory gains in legitimacy as it tends to satisfy or capture considered moral judgements, or approximate whatever the shape of common-sense morality ostensibly is for a given *Umwelt*.

The problem of legitimacy then points to the troubling fact that many moral theories (and especially of the rationalist, maximizing variety) do not approximate common-sense morality in this way, and instead, tend to *revise* upon it. Utilitarianism, for instance, is concerned only with the impersonal maximization of expected utility, and any other potential normatively relevant factor is either subsumed and underdetermined by this calculus, or cast away as fallacious[9]. Rawlsian ethics, at least in its treatment in the machine ethics literature, is concerned only with maximizing the lot of the least well-off agent in the AMA's environment, rather than additionally assessing *who* this agent is to the AMA, or whether he (or the AMA) is breaking any moral rules, and how this all aligns with utility. This will to revise, we will claim, is a methodological carry over from the field of moral philosophy which has dubious foundations in the field of machine ethics. Although the design of artificial morality may have the *beneficial side effect* of revealing moral tenets and complexities which were hitherto either inaccessible to humans, or deeply tacit in their behavior, it is another thing entirely to maintain that the purpose of artificial morality itself is to accomplish this theoretical progress.

[6] Such a line has been followed by (Bharghava & Kim, 2017) and (Talbot et al., 2017).
[7] This type of application is likely championed by Ronald Arkin's work on lethal autonomous weapons (Arkin, 2009).
[8] We will omit the fourth option of justifying the implementation of a given moral theory given an AMA's failure to satisfy one or more of the conditions of the standard view of moral agency, as is done for instance in Talbot and colleagues' (2017) espousal of utilitarianism.
[9] Kagan, 1991.

Instead, this revisionist quality of many maximalist accounts of artificial morality forces us to choose *which* considered moral judgement is to prevail in an AMA's moral responsiveness, rather than attempting to accommodate as many such judgements as possible, and in so doing, ironically approximate the minimalist's ideal. This we will argue shortly, leads to a form of disappointment which is antithetical to the purpose of artificial morality. It seems plausible, in other words, that the real purpose of artificial morality is to approximate acceptable behavior while avoiding unethical consequences, rather than to hold that an AMA should murder Aunt Agatha, sacrifice its passenger, or give its user's groceries away without consent, all in the name of the stringent application of an arbitrary moral principle which allegedly leads to the 'best' or 'correct' results. Put simply, it seems questionable whether the mobilization of any moral principle, however laudable, warrants a robot's *under*-responsiveness to the moral expectations of human agents.

With our position in place then, this chapter is laid out accordingly: In section one, we seek some justification for the minimalist's ideal in the origins of artificial morality. This we lead us to reevaluate what we have called the *argument from increasing automation*, and the ultimate purpose of artificial morality. In section two, we attempt to reconcile a maximalist approach with the minimalist's ideal through the elaboration of a theoretical decision procedure, which we will call the *Ethical Valence Theory*. In this chapter, only a theoretical introduction is provided, while a more contextualized treatment is given in chapter VII, through the application case of autonomous vehicles.

1. *Expanding on the Argument from Increasing Automation*

Recalling from chapter 2, we maintained that the argument from increasing automation (AFIA) served as the principal justification for the implementation of artificial morality into artificial agents, or the justification for the shift from artificial agent, to artificial *moral agent*. In detail, AFIA underpins the idea that if moral value is an inherent feature of an artificial agent's environment, then a failure to account for this moral value in the agent's decision-making will result in the risk of human harm. The more autonomous an artificial agent is, the more this risk is

increased. At this point in our analysis, it should be pertinent to revisit a number of this argument's key terms, so as to better align its repercussions with some points made previously.

First, we must qualify what is meant by our scalar use of the term 'autonomy'. In chapter two, we maintained that there were at least two senses of the engineer's concept of autonomy: provision and independence, where these concepts tracked (loosely) a top-down or a bottom-up approach to the design of an agent program. Further, we addressed a type of continuum of machine autonomy, which was characterized by an inverse relationship between machine autonomy and human control or oversight. In this sense, it was as much a matter of contextual implementation as of programming style that determined the autonomy of an AMA, since a provision-type AMA might still have sweeping independence over a wide range of action and decision contexts, while it was in many respects the fear of *decisional opacity* rather than autonomy *stricto sensu* which stoked much of the fire of the moral concern for autonomy as independence. Given the general trend of our argument, it should now be clear that the type of autonomy with which we are here concerned is of the *provision* variety, pertaining to top-down systems, or expert programs. Accordingly, if the principles and procedures that move the agent from perception to action in its environment are no small mystery, it is precisely the scope of the agent's practical agency—*where* it applies these principles and procedures—that tracks this scalar sense of autonomy. This use of autonomy then aligns with what we have been calling ethically salient contexts, where the more an agent is granted autonomy—and therefore freedom from direct human input or oversight—in situations where moral value is an inherent feature of an environment, the more pressing the need for artificial morality becomes[10].

Second, it seems natural to address what is meant by the idea of moral value being an inherent feature of an environment. Tacitly, our paying attention to the idea of explicitly *human* harm in AFIA paints a rather restrictive and traditional picture of what this could entail. In this restricted sense, moral value seems to pertain to creatures or objects with, for instance, a high degree of organic unity[11], such as human agents and certain sentient animals, or those that make it

[10] In a broader sense, this idea aligns well with what is often called 'Moor's Law' in the machine ethics literature: "As technological revolutions increase their social impact, ethical problems increase" (Moor, 2005, 117).
[11] Nozick, 1981.

into the 'club of constituents' of a given moral theory. The unifying feature of this class of entities is the *intrinsic* quality of their value: they are not instrumentally valuable, for instance in the pursuit of human ends or interests, but are seen to be bearers of value in and of themselves. However, depending on the type of *Umwelt*, the class of bearers of moral value can likely be extended in more practical ways. For instance, a search and rescue robot operating in the Canadian Rocky Mountains would surely consider human agents as subjects of moral value, but perhaps also certain endangered species such as the Bald Eagle or the Grizzly Bear.

A general discussion of what types of objects and entities can be considered to bear intrinsic value seems somewhat orthogonal to our project here, since an exhaustive account of the bearers of intrinsic value only seems to be required for universal practical agents who must distinguish these entities everywhere. Instead, we can make the relatively indisputable claim that human agents themselves possess moral value (flowing from their status as a moral agent, or as a 'value-seeking I'), and that certain entities, depending on an AMA's *Umwelt*, may also benefit from this status. Critically, the idea of moral value is inextricably linked with the possession of moral patiency and thus a *moral push*, which is to say that these entities are seen to hold moral claims over the AMA's action, and that these claims together make up the 'moral pull' to which the AMA is seen to respond. In this sense, where there is moral value, there is moral pull, and where there is moral pull, there is a need for artificial morality.

Finally, however, it seems questionable whether the presence of human agents, together with a significant degree of machine autonomy alone are sufficient to constitute what is meant by an ethically salient context, and thus the setting for which the argument from increasing automation is meant to apply. For instance, it seems somewhat strange to hold that a robot vacuum requires an artificial morality (even perhaps a highly 'autonomous' fictitious one which would say, vacuum, mop and dust), even if human agents (and perhaps pets) are likely often present in its environment. Similarly, if the SoupSaint2020™ had an unlimited supply of soup to hand out to homeless individuals, it is not entirely clear whether it would require an artificial morality in order to inform an optimal distribution of soup. Here, we will claim that the intuitions behind these examples point to a necessary, if not conventional feature of what most machine ethicists consider to be constitutive of an ethically salient context: a conflict in moral claims. This is to say that wherever

an AMA is unable to satisfy all of the moral claims which could arise as a result of its practical agency, a choice concerning *which* claims to satisfy and which claims to reject must be made, and artificial morality constitutes the agent program which manages these sorts of choices[12]. From a structural angle, it would then seem that the design of artificial morality constitutes an exercise in *moral claim mitigation.*

We must pause here to take stock of an important ramification of this claim. In effect, this particular perspective can shed much light on what is meant by the final piece of AFIA, the 'risk of human harm', particularly because it prevents the argument from increasing automation from collapsing into a pure appeal for machine *safety*. This type of deflationist argument has recently been levelled by Aimee van Wynsberghe and Scott Robins, in their explosive article, 'Critiquing the Reasons for Making Artificial Moral Agents'[13]. Their argument proceeds in three steps, which we will do well to lay out here. First, they identify a connection between the implementation of AMAs and a reduction in human harm:

> For many scholars the development of moral machines is aimed at preventing a robot from hurting human beings. To ensure that humans can overcome the potential for physical harm, a technological solution is presented; namely, to develop AMAs…This also speaks to the interconnection of the reasons in favor of AMAs; robots are inevitable, robots could harm us, therefore robots should be made into AMAs[14].

[12] At first blush, this might appear to be a rather restrictive position. However, as our discussion of the *place* of artificial morality in chapter IV explored, the types of facts and features which are seen to underpin a claim can vary widely, and with this variance, comes significant opportunity for conflict. To illustrate briefly, even the most permissive views of what can count as an ethically salient context can be seen to mobilize some sense of claim conflict. Mattias Scheutz' notion of a 'morally charged context' (2016) is probably one of the most permissive in the literature (van Wynsberghe & Robbins, 2019), counting activities such as scaring cats and misunderstanding vocal commands as instances of harm (2016, 3). Still, his view unites all of these cases under the common banner of "…ordinary life decision-making situations in which multiple agents are involved and where a decision maker's available actions can impact other agents in different ways, causing harm to some while sparing others and vice versa depending on the circumstances"(2016, 7). Here too, it would seem that the activity of sparing and harming, or what we might more reasonably call 'benefiting' and 'failing to benefit' could count as an instantiation of competing claims. In this sense, it is the tension across competing claims, rather than the substantive definition of a claim, which is a necessary condition of an ethically salient context.

[13] van Wynsberghe & Robbins, 2019.

[14] Ibid., p. 725.

They then equate the concern for human wellbeing in the context of robotics with the more conventional concern for human safety in technology development:

> There are plenty of technologies capable of harming human beings (e.g. lawn mowers, automatic doors, curling irons, blenders); the solution has always been either to design them with safety features or to limit the contexts in which a technology can be used[15].

Finally, they maintain that the equation of morality with safety speaks to an impoverished understanding of the conceptual territory of the former, and conclude from this that the use of 'morality' in the AMA debate constitutes what they call a linguistic 'trojan horse':

> …the real concern for ethicists is that ethics is being reduced to safety. Notions such as values, rights, freedoms, good vs bad, right vs wrong, are central to the study of ethics…One may believe that the values of safety and security are fundamental to achieving the good life; however ethics cannot be reduced to these issues. So if AMAs are simply a solution to possibly harmful machines, then *safety*—not *moral agency*—is the object of debate…the word 'moral' is a linguistic trojan horse—a word 'that smuggles in a rich interconnected web of human concepts that are not part of a computer system or how it operates'[16].

If van Wynsberghe and Robbins' argument as to this moral 'trojan horse' is convincing at all, it is likely due to their surreptitious precision in step one that it is indeed *physical harm* that machine ethicists seek to prevent in their appeal to artificial morality, a precision which is missing in many accounts of the argument from increasing automation[17]. If this were the case, it would indeed seem that the simulation of moral agency in AMAs is as inappropriate as it would be in lawnmowers.

---

[15] Ibid., p. 725.
[16] Ibid, p. 725: Sharkey, 2012, 793.
[17] In fact, the two examples the authors themselves use in this argument do not restrict their concept of harm to purely physical states: "the only way to minimize human harm is to build moral competent robots that can detect and resolve morally charged situations in human-like ways" (Scheutz, 2016), "it is clear that machines…will be capable of causing harm to human beings" (Anderson & Anderson, 2010). This particular point was also briefly addressed in Ben Byford's contribution to Poulsen et al.'s (2019) rebuttal of this article: "As both a researcher and technologist one is often faced with both the philosophical and the practical. In the cited case I agree with the argument that 'harm' should be categorized to include other forms of indirect harm, not only physical" (2019, 7).

However, it seems highly questionable whether the conceptual extension of the values of safety and security—and the likely maximization thereof—totally covers the types of moral responsiveness most machine ethicists seek to simulate in machines. In this sense, a robot care assistant that must balance the desire of a patient to refuse medication with its duty to adhere to the Hippocratic oath may be open to considerations of physical harm, but these considerations do not seem easily reducible to a concern for safety or security. Moreover, the *general* design of both a mountain search and rescue robot, and SoupSaint2020™ may involve some degree of safety considerations, however the decision procedure that decides who to save or who to feed when not everyone can be satisfied doesn't uniquely track the maximization of user safety. Finally, it would seem that the causing of (even indirect forms of) physical harm entertains an odd relationship with *virtual agents*, even though the behavior of these has certainly led many a machine ethicist to pronounce an argument from increasing automation.

It seems then that if there is a linguistic trojan horse hidden within the argument from increasing automation, it is to be found in the concept of 'harm', rather than the concept of 'moral'. Indeed, maintaining that physical harm is a *morally relevant feature* of many ethically salient contexts is certainly defensible and likely common, but to deduce from this that the entire enterprise of machine ethics is reducible to the minimization of this feature in robotic action seems to miss the mark. There are certainly other normative constructs that underpin human expectations in ethically salient contexts, pertaining precisely to "values, rights, freedoms", and particular views of "good vs bad, right vs wrong". By consequence, it should be clear that human agents are able to be harmed in more than just physical ways, and it is precisely safety from *this* type of harm that the argument from increasing automation, and the concept of moral responsiveness seeks to capture.

Indeed, while van Wynsberghe and Robbins were surely right to point out the prevalence of harm-based accounts of the need for artificial moral agents in machine ethics literature, there exists another rather popular design goal which better expounds the type of harm machine ethicists seek to avoid. This relates to the goal of designing a *praiseworthy* artificial moral agent[18]. Initially,

[18] Allen, Varner & Zinser, 2000: Wallach & Allen, 2008.

the concept of praiseworthiness does not seem to readily apply to explicit ethical agents, owing mainly to their incapacity for autonomous normative endorsement, or their inability to be motivated morally[19]. Here again in other words, the fact that artificial moral agents fail to satisfy the conditions of the standard view of moral agency seems to remove them from any possible entry into the constituency of praiseworthy agents. As one author maintains:

> Moral praiseworthiness, at least in contemporary moral practice, is closely linked to the capacity to make autonomous decisions, particularly in weighing moral reasons against pressing reasons of other sorts—self-interest or the interest of those one cares for, private political commitments, and so forth. Moral praiseworthiness does not apply to an agent who has literally no other choice than to follow the moral rules that have been programmed to override all other reasons in contexts of conflicting reasons[20].

Since current forms of AMAs are unable to weigh reasons of these kinds against one another in their practical deliberation[21], and since the principles upon which they act are not autonomously chosen, the necessary *internal* conditions of praiseworthiness are not met—AMAs are not able to act on a moral reason simply in virtue of its being a moral reason. Here again in other words, an AMA's failure to meet the conditions of the standard view of moral agency serves as the crux of this critique. It is certainly plausible however, that the use of the term 'praiseworthiness' in the context of AMA design is not meant to denote this conceptually thick, anthropomorphic capacity, but rather something more shallow that still avoids falling into total vacuity. To this end, Wallach & Allen, to whom the first use of this term can be attributed, seem to focus on the *external* perception of the AMA as a praiseworthy agent, precisely in its capacity to act for the right or appropriate reason:

[19] A moral motivation, according to Bernard Williams, consists in "…motivations that spring from thinking that a certain course of action is one that one ought to take" (1976, 174). Williams maintains that in ethical theory, these are typically dissociated from what he calls *natural* motivations, that may manifest themselves in feelings such as regret or distress over acting in ways which thwart these moral motivations (1976, 174-175). While Williams sees this disassociation as a shortcoming of many moral theories, it stands to reason that level 4 explicit ethical agents do not have either type of motivation, regardless of their connection.
[20] Podschwadek, 2017, 337.
[21] Or on an even stricter view, are not able to be *responsive* to many of these reasons (Purves et al., 2015).

> The discordant and sometimes obviously anthropocentric theories of the various ethical schools do not bring one very close to clear criteria for…treating an autonomous (ro)bot as a moral agent. However, one should not conclude that systems incapable of comprehending the effects of their actions will not be morally praised or blamed for these effects. Human tendencies to assign praise and blame are complex and subject to many influences, and there is every chance that they will be extended to (ro)bots. An AMA might be considered *praiseworthy* once it has the capacity to assess the effects of its actions and to use those assessments to make appropriate choices[22].

Here then, the concept of a praiseworthy artificial moral agent seems to be tethered to the idea of *external approval*: whatever the inner workings of the robot are, the outward behavior it displays might constitute praiseworthy behavior if it can be seen to align with the types of normative or moral standards that guide human behavior in similar contexts. It is praiseworthy in other words, as soon as the human agents with which it interacts find its decisions appropriate or acceptable. In this sense, the job of identifying, interpreting and endorsing these standards is punted back to the designer of the AMA in his choice of whichever criterions or principles inform the AMA's artificial morality. There must be familiarity in output, but there can be significant dissimilarity in the ways by which this output is achieved.

Returning to the concept of claim mitigation, we now appear to have a clearer idea of how this concept is linked with the risk of human harm. In effect, the idea that artificial morality tracks something more than the prevention of physical harm to humans seems to translate to the idea that the *moral claims* that individuals hold over the AMA's behavior are not *uniquely* underdetermined by facts about individual physical well-being. Put another way, there is more to the moral push of a human agent than a simple moral injunction to refrain from physically harming her, indeed there seems to be multiple ways that an AMA can be responsive to her moral value. We will flesh this claim out with a few of our previously cited examples.

Let us begin with the case of the robot personal shopper from chapter IV. Let us assume momentarily that this shopping assistant is human, perhaps a neighbor of the original user. This

[22] Wallach & Allen, 2008, 201.

neighbor is sent to a store to buy, say, apples for this user, presumably using the user's funds. When she encounters the homeless person on the way back from the store, what sorts of considerations likely figure in her decision to give some of the groceries to this individual? Ostensibly, the neighbor must consider two claims in her decision: the claim of the original user to the groceries he has paid for, and the claim the homeless person holds as to badly needed sustenance. Other considerations might also seem pertinent in her deliberation: the potential interest or preferences of the original user, the fact that the homeless person is in greater need of the apple than this user, and perhaps the fact that the tacit agreement between her and the user did not admit any conditions as to charitable donations. In this sense, each consideration provides a certain weight to the claim of either the user or the homeless person, and the neighbor must weigh these claims in order to decide whether to give the apples to the homeless person, and if so, how many.

If we apply the same experiment to the case of the robotic cashier, a similar structure of competing claims emerges. Again, the cashier seems torn between two claims: the claim the shop owner holds over the rightful sale of his commodities, and the claim of the customer to purchase whichever goods he pleases. Further considerations may also help the cashier weigh these claims: the fact that the sale of cigarettes is legal, the fact that the cashier is contractually obliged to sell the goods in the store, and the fact that the customer is a sober, consenting individual of legal age, who is well informed about the negative effects cigarette smoking can have on long-term health. Finally, in the case of the aunt Agatha terminator, the prospective murderer must weigh two different types of claims in his decision to end Agatha's life: the claims of the individuals who are likely to benefit from aunt Agatha's massive fortune if her life were to abruptly end, and the claim Agatha holds as to her continued existence. The consideration that her fortune would be squandered on feral cat sanctuaries if her life were to continue, as well as the fact that many lives could be improved as a result of her fortune going to charity, again appear to give weight to these competing claims.

First, it should be obvious that the type of claims that are entertained in these processes of deliberation flow from different types of moral value, moral rules or moral ideals. The Aunt Agatha Terminator appears to be caught in a conflict between a moral injunction to refrain from intentional

harm (an instantiation of the 'harm principle'[23]), and the impersonal claim to utility that her timely death would satisfy. The cashier seems to be caught in a slightly different snare between refraining from harming a customer, and respecting both that customer's autonomy and potentially the contractual obligations of employment she is required to uphold. Finally, the personal shopper again seems to be caught between a special obligation to respect the original user's interest (flowing perhaps from a promise, or from her acting as a proxy for this user), and the claims to welfare that a starving homeless individual exacts on her.

Second, it should be clear that each type of claim provides the decision-maker with a *pro tanto* moral reason for action, or in stronger language, each claim considered separately translates to a moral requirement for the agent to act in a certain way. In this sense, a customer's claim to individual autonomy provides the cashier with a reason to respect it and thus to sell her whichever goods she pleases, just as a consideration of this same customer's welfare seems to provide a reason to refrain from selling her harmful goods. Finally and critically, a failure on the part of the decision maker to respect the claims of these individuals seems to cause a harmed-state[24] in them, flowing less from a concern for their immediate physical welfare, and more from a *non-respect* or insensitivity to the value of this individual. In this way, ignoring a homeless individual's claim to sustenance constitutes a form of harm to this individual, which may not exclusively manifest itself in an actual reduction in individual physical wellbeing, but rather in the form of a legitimate *complaint* that the deliberation of the personal shopper did not adequately take into consideration the moral pull that the homeless individual's moral value exerts on her decision: it failed to act in ways which garnered *external approval*. As Bernard Williams maintains, "The notion of a moral claim is of something that I may not ignore: hence, it is not up to me to give myself a life free from conflict by withdrawing my interest from such claims"[25]. Put succinctly, unresponsiveness or perhaps under-responsiveness to the moral claims (or moral pull) of individuals constitute a broader, and morally salient sense of the concept of harm, one which tracks external approval, and to which a praiseworthy AMA is necessarily sensitive.

[23] *Sed iustitiae primum munus est, ut ne cui quis noceat, nisi lacessitus injuria* [the first property of justice is, that no one should harm another, unless provoked by a prejudice]. (Cicero, 1913, 22).
[24] Feinberg, 1987.
[25] Williams, 1976, 178.

Accordingly, let us interpret for a final time the impact this discussion has on the argument from increasing automation. If it can be maintained that the concept of harm tracks something more than purely physical welfare in human beings, then it would seem that AFIA seeks to implement artificial morality as a way to *maximize the artificial moral agent's moral responsiveness to the value of the individuals it encounters*. More precisely, the role of artificial morality seems to be to maximally satisfy the moral claims exerted on its behavior by the bearers of moral value in its *Umwelt,* since this would lead to the greatest reduction in the risk of human harm, understood in this special sense. Artificial moral agents are then akin to *maximal moral pull responders*, selfless agents whose decisions are maximally constrained by the moral value they encounter as a result of their practical agency.

If we were then to translate this idea in the form of a design principle for artificial morality, it would appear to resemble what we might call *the principle of total irreproachability* (PTI): an artificial morality must be so designed as to ensure a maximal degree of responsiveness to the moral pull of the agent's *Umwelt,* or to the moral claims exerted on the agent by bearers of moral value. Evaluated negatively, the PTI then seems to forbid any insensitivity to an individual's moral claim, since this would result in the causing of harm to a bearer of moral value. The next section explores the cogency of this principle, since as we might suspect, it proves particularly challenging to satisfy in the design of many artificial moral agents.

1.1 *Exploring the Principle of Total Irreproachability in the Design of Artificial Morality*

To begin our exploration of the principle of total irreproachability, it might be useful to revisit a concept which we briefly addressed in chapter 4, that of *normative convergence*. This concept was meant to capture the idea that consensus concerning the normative standards of moral behavior in AMAs was both decipherable and attainable in many *Umwelts*; in stark contrast to more *dilemmatic* contexts in which rational and informed individuals could still be seen to disagree about what morality required from AMAs[26]. To recall, this concept implied either a narrow or

[26] Tolmeijer et al., 2020.

moderate view of the place of artificial morality, and was supported by Dignum's second criterion for the design of top-down approaches to artificial morality.

In one sense, the idea that human agents are able to agree on what types of normative standards robots should adhere to is intimately linked to the types of *roles* they perform in society. In philosophical literature, it is no small secret that many human agents must endorse specific principles or frameworks when serving in institutional roles such as doctors, lawyers, or politicians; and that these standards are, in Bernard Williams' words, 'logically welded to the title'[27]. In this sense, the ethical constraints endemic to the discipline of medicine constrain the behavior of doctors, and all doctors are likely judged in terms of their adherence to these standards. Thus, when designing an artificial agent which will act within the *Umwelt* of a hospital for instance, it seems plausible that these same standards could inform the types of desirable behavior that its agent program, and perhaps eventually its artificial morality, should exhibit. Further, it seems plausible that an 'optionless' application of these standards is desirable, thus painting these machines as pure distributive agents.

However, it should be equally clear that many robots, in virtue of their complex purpose-oriented ontology, serve somewhat unprecedented roles—or at least, serve roles which are not entirely interchangeable with those of the humans who once carried out their functions. In this way, an autonomous vehicle seems to occupy a space somewhere between a taxi driver, private chauffeur and a trusted or designated driver, while a home healthcare assistant seems to float somewhere between a caretaker, a diagnostician, a nurse and a personal companion. In these types of cases, it would seem that the type of standards of assessment which would lead to normative convergence are less clear or mostly implicit, or in any case, may spring from more than one sort of professional ethical doctrine or established practice. In this sense, general epistemic uncertainty regarding how to program these robots is likely greater, while still remaining somewhat clearer

[27] "…various sorts of title or role can conceptually carry with them broad standards of assessment of people under those titles, as the descriptions of artefacts can carry standards of assessment of those artefacts. While the standards can be in this way logically welded to the title, the title is not logically welded to the man; hence, the standards are not logically welded to the man" (Williams, 2012, 52). See also (Johnson & Powers, 2008).

than the entirely *unprecedented* decision contexts which some AMAs may encounter—such as an autonomous vehicle's *deliberative* decision as to 'how to crash' in unavoidable collision[28].

Thus, at this general level, it would seem that it is easier to decipher what morality requires—and what society may correspondingly expect—of some types of robots, while others are seen to disrupt the fabric of normative standards in more incisive ways. To usher in some of the vocabulary from the previous section, we can further substantiate this idea by maintaining that in situations where broad normative standards are clearly established, there exists some type of agreement concerning *how* the robot ought to manage the different claims it encounters, yielding a specific mitigation strategy which weighs and balances these. A specific claim mitigation strategy, in other words, comes part and parcel with the institutional role a robot is designed to uphold. In this sense, the concept of normative convergence may extend into even the more 'dilemmatic' corners of fields such as healthcare robotics, in virtue of a) the fact that these decision contexts have long been encountered by human agents—and sometimes on a daily basis[29], and b) stringent and precise decisional procedures, principles and rules have been elaborated to manage the types of conflicting claims that may occur in this *Umwelt*[30]. Accordingly, even if the claims of some human agents are dismissed in some cases—such as when a doctor must decide which of two patients ought to receive a badly needed liver transplant—the criterion of rightness by which he makes his decision (for instance, maximize life years, or the expected success of the transplant) may nevertheless be *acceptable*, or garner external approval from all parties, and thus less likely to engender the type of complaint indicative of our concept of harm.

However, this leads us to the corresponding idea that for those robots whose roles are unprecedented or under-defined, normative consensus surrounding how they ought to resolve claim conflict can be similarly lacking. In this sense, it seems plausible that it is not only the

[28] Lin, 2016.
[29] van Wynsberghe & Robbins, 2019.
[30] For instance, the United States Department of Health and Human Services Advisory Committee on Organ Transplantation recommendations establish a list of admissible features by which a patient is to receive priority for a transplant. These include option features related to the potential success of the transplant (blood type, body size), context features such as the distance between the donor's hospital and the patient's hospital, and relational features such as the severity of the patient's medical condition or his waiting time. The full list is available here: https://www.organdonor.gov/about/process/matching.html.

designers of these AMAs who operate in an environment of epistemic or moral uncertainty as to how to resolve conflicting claims, but also, that its human users and perhaps society at large are also at a loss for what to *expect* in terms of responsiveness to their claims. To put this in clearer terms, we might hold that in these cases there exists significant epistemic uncertainty regarding which mitigation strategies are *acceptable* to society, and correspondingly, that the choice of any particular strategy seems more likely to engender this special kind of complaint which indicates harm.

Thus, from the vantage point of normative convergence, we are able to better ascertain what the principle of total irreproachability implies for the design of artificial morality. Mainly, we will notice the importance of the connection between this principle and the notion of a 'complaint', and the particular normative status of the latter. In effect, the PTI entertains a complementary relationship with normative convergence: in areas where normative standards are clear and established, so too is the potential form that a robot's moral responsiveness should take. In this sense, any artificial morality which emulates this form of moral responsiveness is less likely to generate the types of complaint of non-responsiveness that the PTI seeks to minimize. This is not a function of the particular artificial morality's espousal of the 'right' or 'correct' criterion of rightness or decision procedure considered *sub specie aeternitatis*, but instead a function of its being *acceptable* in light of the ethical principles or standards that have historically driven the performance of that specific role or institution. Here then, we must consider the idea that while complaints considered generally track non-responsiveness to a moral claim, what we might loosely call a *legitimate complaint* tracks a form of non-responsiveness that also fails to conform to normative convergence, and is for this reason unacceptable. In this sense, if a patient's reception of a badly needed liver transplant is usurped by, say, the president of the United States after an assassination attempt, the patient holds a general type of complaint in virtue of the fact that a hospital's decision to give this liver to the president fails to adequately consider his value as an individual; but additionally, he may hold a 'legitimate' complaint which is grounded in this decision's failure to expound the *accepted ethical practices* endemic to organ donation. The patient expects that his life matters just as much as any other potential organ recipient, and that the decisions as to who receives a transplant will be made in accordance with fair, objective and explicit principles which apply equally to everyone.

Consequently, we can hold that AMAs who are active in contexts of normative convergence are more likely to satisfy the principle of total irreproachability, precisely because well-established normative standards serve to minimize the types of legitimate complaints that human agents may level against the decisions of the machine. This leaves us, however, with the troubling case of those AMAs whose *Umwelts* do not benefit from this type of normative agreement, or whose roles are less easily identified. Indeed, it would seem that the absence of normative convergence erases the useful distinction between general and legitimate complaints, and in so doing, poses a serious challenge to the satisfaction of the PTI. This is because, superficially at least, in the absence of knowledge as to what is an acceptable strategy for claim mitigation, every instance of non-responsiveness to moral claims seems to generate a legitimate complaint. In other words, if there are no general standards to which designers can adhere and to which users can expect the robot to adhere, it is ostensibly anybody's guess what ideal moral behavior resembles, and thus there is no established reason for an AMA to dismiss some claims in pursuit of others in situations where not all claims can be satisfied[31].

Suddenly then, the maximalist's project of implementing the 'right' moral theory again appears attractive, since if nothing else, it provides one set of standards (or a criterion of rightness) which is explicit, and which can serve to justify the AMA's decisions against complaints of non-responsiveness that human agents could hold. In this sense, if it were possible to establish, as a sort of institutional fact or practice, that all robots were programmed to act as pure utilitarians in contexts of claim conflict, then this would ostensibly relegate Aunt Agatha's claim to life to the status of a non-legitimate complaint, and thus justify the Terminator's decision to kill her so as to give all her money to charity. Less contentiously perhaps, if it were possible to establish that all personal shopper robots were to act in adherence to the maximin principle, then the robot's user would not hold a legitimate complaint when some of her groceries are given to a homeless person.

[31] "Whereas there are objective and measurable criteria for the evaluation of thermostats, things are more cumbersome when it comes to moral machines…the complication arises from the question of what exactly is to count as executing the task at hand in morally appropriate ways, or against what exactly the behavior of the system should be evaluated…The ontological and epistemic complications that arise in the moral domain thus make it difficult to settle on standards…More fundamentally, it is not even evident what kinds of considerations should guide the process of choosing such standards" (Tolmeijer et al., 2020, 10).

It is exactly here, in other words, that the 'tie breaking' function of a moral theory appears as a knock-down argument for the conscientious maximalist: places where accepted practices and standards give no recommendation as to what morality requires of AMAs. The use of the moral theory allows us to dismiss certain categories of claims as irrelevant or unfounded, and in so doing singles out only one claim, or one type of claim that the AMA is morally required to respect[32]. In other words, the problem of legitimacy that plagues the maximalist, or the fact that he fails to satisfy the principle of total irreproachability by revising upon common-sense morality, seems to be inevitable. This is so since even if the *choice* of moral theory is arbitrary, its *role* in artificial morality is not: it minimally enables an AMA to exhibit some form of moral responsiveness in the face of great epistemic uncertainty, and maximally might ensure that this resulting responsiveness tracks some vision of the better angels of our nature[33].

Surely however, a lack of normative convergence surrounding some types of AMA does not necessarily lead to the vindication of the maximalist project in such a straightforward way; if only because a great deal of research in machine ethics is dedicated to the discovery of precisely these standards, in the form of empirical research into acceptability as moral preference, adoptability, and institutional viability. The force of the maximalist's position as a *pis aller* in artificial morality only holds if there are absolutely no normative standards which AMAs could reasonably be expected to expound. In effect, asking survey participants what an autonomous vehicle should do in the case of an unavoidable accident counts rather clearly as an exercise in normative convergence development, where researchers attempt to identify exactly which types of claims and complaints can be held to be legitimate in the eyes of users and of society, views which might then be incorporated into a larger iterative process of stakeholder consensus-building. Though these standards are less clear and admit often of less internal coherence than the more established practices of fields such as medicine and law, they still indicate the contours of

[32] "The evident fact that there is at most one of the two things which, all things considered, I should do, is taken to be equivalent to the idea that, all things considered, there is only one obligation. But this is a mistake: There are certainly two obligations in a real case of this kind, though one may outweigh the other. The one that outweighs has greater stringency, but the one that is outweighed also possess some stringency…" (Williams, 1981, 73).
[33] Bhargava & Kim, 2017.

acceptable claim mitigation strategies in the face of epistemic uncertainty. Thus, dismissing the utility of the minimalist's ideal in the outright seems unfounded.

If an even remotely coherent account of normative convergence is available as a resource to artificial morality design, then an agent program which incorporates this data is still more likely to respect the principle of total irreproachability than a moral theory which only views some types of moral claims as true moral requirements, and thus dismisses many individual claims[34]. This strategy, in other words, is more likely to lead to outwardly praiseworthy artificial moral agents[35]. From this vantage point, it seems then that the maximalist is guilty of cherry picking, since if normative convergence uncontroversially guides the design of artificial morality in some *Umwelts* such as medicine, law and warfare, it is unclear why this same property should have no bearing whatsoever in others.

Thus, what is called for in cases where robots serve unprecedented roles, or act in unprecedented contexts, is likely not absolute adherence to some arbitrary moral theory. Instead, what seems just as viable, and certainly more acceptable, is to align the principle of total irreproachability with the types of legitimate complaints that flow from acceptability research. While it may certainly be frustrating that this project likely does not pass through the strict implementation of a moral theory, or for that matter, does not much resemble an exercise in moral theory at all, this does not logically entail the abandonment of all moral prescriptions and recommendations and the pursuit of purely descriptive (and potentially immoral behavior). Instead, the structure of moral theory must be made to cooperate with the burgeoning normative convergence that flows from acceptability studies, amplifying or extrapolating rather than revising

[34] "It seems to me a fundamental criticism of many ethical theories that their accounts of moral conflict and its resolution…eliminate from the scene the *ought* that is not acted upon. A structure appropriate to conflicts of belief is projected onto the moral case; one by which the conflict is basically adventitious, and a resolution of it disembarrasses one of a mistaken view which for a while confused the situation" (Williams, 1976, 175).

[35] In a somewhat distant sense, our claim is supported by the idea that artificial agents are often perceived as social entities by their users—a phenomenon known as *media equation* (Reeves & Nass, 1996: de Graaf, 2016). While there is some empirical evidence for robots being subject to different, and perhaps more stringent behavioral standards than humans serving comparable roles (Malle et al., 2015), it stands to reason that at least in dyadic interactions, the norms that govern human-to-human interaction might therefore resemble the normative expectations humans hold towards artificial moral agents.

upon the types of principles and procedures it recommends. In the following section, we will attempt to delineate the general structure of an artificial morality that accomplishes this marriage between the virtues of the maximalist, the minimalist's ideal, and the principle of total irreproachability.

2. *The Ethical Valence Theory*

Until very recently, the use of the human likeness as a basis for machine design has been a pervasive assumption in the fields of machine ethics and human-robot interaction. In other words, the assumption has been that the probability of a machine's successful performance in a human social sphere increases as it is able to mimic the social capacities of human agents: emotions, reciprocity, empathy, morality, and many more. In practice, this assumption generates what Aimee van Wynsberghe sees as a naturalistic fallacy in robotic design[36]. For instance, if a human is able to deceive another human across interaction, then a robot *ought* to be programmed to deceive; whether as a design end in itself, or as an instrumental means to achieving some other design value such as user acceptance, adoption or trust[37]. What we might call the *mimetic fallacy* then describes the simple design credo, 'what works for humans, ought to work for robots, or for robotic acceptance'.

If we transfer this idea into the field of expressly artificial *moral* agents, then we might see how such a mimetic fallacy has driven the popularity of both top-down and bottom-up models, both of which seek to emulate significant aspects of human processes of moral reasoning. The assumption then, is that if humans can be seen to exercise their moral agency in some decision contexts, then it would seem that robots ought to exercise a similar capacity in these contexts, to a maximal degree of similitude[38]. Interestingly, this tendency is orthogonal to, while not being quite

[36] van Wynsberghe, 2020.
[37] "Robots embedded with social interaction features, such as familiar humanlike gestures or facial expressions in their designs, are likely to further encourage people to interact socially with those robots in a fundamentally unique way. For human users, interaction with robots is, in a sense, more as if one is interacting with an animal or another person rather than interacting with a technology" (de Graaf, 2016, 592).
[38] "Machine ethics could be defined as designing machines that do things which, when done by humans, are criterial of the possession of 'ethical status' in those humans" (Torrance, 2011, 118).

antithetical to, the idea of morality being mobilized so as to reduce the harmful impact of AMA behavior. This is so since, while harm prevention and even moral responsiveness figure prominently in the spectrum of human moral behavior, they together hardly constitute the full picture: leaving out most prominently, the moral push and freedom that flow from the status of a moral agent itself. In this sense, it seems neither useful nor desirable to mimic the moral behavior of humans past the point at which total irreproachability is achieved in AMAs, since it is ostensibly at this point that an AMA moves from partner to *other*[39]. Indeed, if it is really this ideal of partnership, rather than artificial *personhood* that we are after, then the best we can hope to design is maximal moral pull responders.

In line with our analysis in chapter IV, it would then seem that the mimetic fallacy is particularly ill-suited to guide the design of artificial moral agents, since patently, half of the human moral picture is necessarily missing in these 'pushless' agents. Put another way, it is only for lack of better options that we seek to mimic human moral reasoning in these machines, since this reasoning will necessarily be capped at a very crucial point, and therefore be fundamentally dissimilar. Nevertheless, what could a decidedly non-mimetic, or *exotic* approach to artificial morality reasonably entail? Immediately, we should dismiss any option which points to the design of a radically *new* morality, which for instance, manages the extremes of blue and orange, rather than good and evil[40]. Surely, what we would instead seek to design is a functional morality that is *complementary* to this full human moral picture, without at the same time threatening to take up space within it. In this sense, it seems reasonable that some of the signposts of the previous section—the minimalist's ideal and the principle of total irreproachability—might provide us with an entirely exotic way to ensure ethically acceptable behavior in AMAs.

[39] Our use of the term 'partner' differs slightly from its standard use in AI ethics literature. In effect, the established category that best describes level 4 explicit ethical agents is likely that of an 'ethical assistant', agents with "…limited autonomy but [who] are aware of the social environment in which they interact. These systems are expected to have functional morality, meaning that responses to ethically relevant features of the environment are hard-wired in the system architecture…including the possibility to decide not to comply with the norm" (Dignum, 2019, 87-88). We have nevertheless chosen partnership as the operative metaphor here, albeit to convey a very general sentiment.

[40] 'Blue-and-Orange Morality' is a term coined in TV fiction to denote a character which confounds the typical behavior of a moral agent, abiding by his own, relatively obscure moral laws. The members of the Addams Family, in this sense, are all blue-and-orange characters.

In this section, we will introduce one such attempt at an exotic artificial morality, which we will call the Ethical Valence Theory (EVT). While this theory deals in human moral principles and preferences, we shall see that it does not arrive at them in mimetic ways. Instead, the EVT should be understood as a theory of *moral claim mitigation*, which in accordance with our previous analysis, seeks to satisfy the principle of total irreproachability in AMAs destined for unprecedented roles, and dilemmatic contexts. In this sense, the EVT presupposes a narrow view of the place of artificial morality, and aims to provide acceptable action-guiding recommendations in decision contexts where not all claims are jointly satisfiable. Moreover, we assume that these claims are underpinned by fairly strong moral requirements, such as a significant risk of human harm, or a significant risk of insensitivity to a value such as personal autonomy[41]. In effect, the fundamental assumption of the EVT is that any and every human individual (or in certain *Umwelts*, each and every 'bearer of moral value') holds a *claim* on an AMA's behavior, as a condition of their existence in that decision context. The goal of the AMA is to maximally satisfy as many claims as possible as it moves through its environment, responding generally in proportion to the strength of each claim, and to the strongest claim when not all can be satisfied.

The theory itself contains three different structural elements: claims, valences and moral profiles, which together provide a specific strategy as to how to mitigate the competing expectations of human agents through the AMA's behavior. Importantly, the Ethical Valence Theory should not be understood as providing a particular *account* of morality—one criterion of rightness or decision procedure, or one conception of the right or the good—but instead, should be seen as a larger structural framework within which many different accounts of morality can be implemented, depending on the types of facts and features on which the claims, valences and profiles are based. In other terms, this implies that the EVT admits what Tolmeijer et al. call a *diversity consideration*:

[41] In detail, the Ethical Valence Theory, both as a computational and theoretical decision procedure, is not necessarily limited to narrow artificial morality. Hypothetically at least, different accounts of what we will come to call 'moral profiles' can be triggered by the emergence of different ethically salient contexts, tracking a more moderate view, or could be consistently active in the agent program of the AMA, in alignment with a wide view. However, for the purposes of explanation and with the will to endorse the most common perspective in machine ethics and its conditions, we will presume a narrow view for the duration of this chapter and the next.

> …the possibility that not all ethical machines adhere to the same ethical theory type, and the [resultant artificial morality] includes the choice of diverse types of ethics to be implemented…It is considered part of the implementation dimension rather than the ethics dimension since diversity considerations can also exist within the same ethical theory, for example, by allowing deontological machines to have different rules to adhere to while still all being deontological in nature.[42]

When defined this way, Tolmeijer et al.'s definition seems to convey two essential claims. First, that a diversity consideration precludes the possibility of a universal machine ethic, and second, that diversity amounts to a type of interchangeability across different normative theories. Metaphorically, this might resemble an AMA's swapping his 'Kantian' moral glasses for a 'Rawlsian' pair, according to what best suits the occasion[43]. In this sense, the criterion of rightness may change, but the loyalty to moral theory perseveres. While this type of diversity can be accommodated by the EVT, it is perhaps not the best use of its flexibility. Instead, what the EVT aims to achieve is a truly *customizable* ethics setting. In this sense, different normatively relevant factors and morally relevant features can be combined in ways that yield entirely novel criteria of rightness, such as 'an action is right if it spares the most vulnerable affected road user, but does not cause lethal harm to the vehicle's passenger'. In this way, the structure of normative theory is split apart, and made to maximally adhere to our considered moral judgements. The diversity consideration then does not reach across established moral theories, but rather across what can be seen to matter morally to various stakeholders.

At this juncture, it might be important to provide some preliminary remarks concerning why this type of stakeholder inclusion is desirable. While it is true that the more philosophical vein of machine ethics has occupied itself with important, albeit theoretical questions concerning the ideal aspects of robotic behavior, it is also true that the design and development of AMAs is a interdisciplinary, practical question with global reach and impact. In this sense, while machine ethicists are primarily concerned with the goal of creating safe, irreproachable agents, other important actors, such as the original equipment manufacturer, may introduce other goals, limits

[42] Tolmeijer et al., 2020, 9-10.
[43] Or, to do true justice to the definition, that the AMA is also able to swap different *deontological* glasses, alternating for instance, 'the law of armed conflict' set with the 'mission specific rules' set.

and incentives which will nevertheless influence the ultimate design of many AMAs. This is not to imply the more defeatist claim that artificial morality does not or will not matter to private industry, but rather, the more pragmatic claim that offering a flexible framework for artificial morality might increase the likelihood of stakeholder approval and involvement itself. In other terms, even if artificial morality addresses moral claims, it will also need to accommodate the legal, financial and industrial claims of those who have invested in its research and development.

To this end, the Ethical Valence Theory is the result of a publicly and privately funded research project, the AVEthics project, which seeks to unite philosophical, psychological and computational expertise under the banner of societal and industrial approval[44], specifically in the development of ethics policies for autonomous vehicles. This is to say that many of the structural choices presented in this chapter are not the pure product of philosophical investigation, but rather the product of an extensive discussion and negotiation between researchers and industry partners. Candidly, this negotiation has not always been simple or straightforward, mainly due to the many hermeneutic gaps which exist between the academic and industrial understanding of terms such as 'autonomy', 'rationality', 'ethics' and 'moral preference'; but also due to the understandable reticence of these actors to pronounce positive statements or policies surrounding ethical questions for which there is never a single, popular and universal answer. In this sense, the experiences of stakeholder participation within the AVEthics project likely reflect the empirical conditions of most collaborative projects with industry involvement, where stringent analytical coherence may run up against industrial feasibility and enthusiasm.

Then, it is perhaps in light of these experiences that the value of diversity is taken so seriously within the Ethical Valence Theory, and why it presents such a peculiar proposal in comparison to other approaches in the literature. The EVT's goal, in other words, is not the thorough application of a given normative approach to a computational decision procedure or to a given type of AMA. This typically yields a more or less viable account of the extensional definition of morally relevant factors and features within the *Umwelt* in question, contributing to a larger theoretical debate about what matters morally. In some sense, the aims of the EVT point to the

[44] Dogan et al., 2016 : Evans et al., 2020.

phase immediately *following* this theoretical exploration, where available knowledge about what matters morally, and how to make it matter to a machine, are organized in a way that reflects burgeoning normative consensus, and then implemented into a commercial product.

Our introduction of the theory will proceed in three steps. In the first section, we will address the most abstract aspects of the theory, including its normative foundations and the type of division of decisional labor it entertains with the human programmer of an AMA. We will do this by linking the Ethical Valence Theory to contractualist thought and principles, however this is a relatively loose association. In actuality, the elaboration of the EVT within the AVEthics project proceeded *a*theoretically, and its resemblance to Scanlonian forms of contractualism was only discovered *a posteriori*, rather than being established as an *a priori* theoretical conviction[45]. Accordingly, we will see in section I that the EVT resembles a pluralist form of act consequentialism which abides by contractualist constraints, but the legitimacy of this approach flows from stakeholder collaboration, and not from its connection with Scanlonian thought.

In section II, we introduce the concept of a *claim* and a *valence*. These two constructs address the morally relevant features of an AMA's *Umwelt*, tracking the demands of morality, and the minimalists ideal, respectively. In this way, what we have called *claim mitigation* consists in the activity of balancing the claims and valences of the individuals affected by an AMA's decision. It is also here that the perverse incentive problem, and the further feature problem, impose a threshold of moral blindness on the precision of these claims and valences, by limiting the types of constitutive and scalar facts which can underpin them. Finally, the satisfaction of the principle of total irreproachability is managed by the concept of a *moral profile*. We address this construct in section III. In effect, moral profiles attempt to capture the mitigation strategies revealed in local common-sense morality, and it is here that various normatively relevant factors are honored in the AMA's decision procedure. A moral profile, in other words, decides what must be *done* with the claims and valences of affected parties: maximization, trade-off, minimization, prioritization, and so forth. In this sense, each moral profile yields its own theoretical decision procedure, and its own criterion of rightness.

[45] Much credit is due to the comments of Geoff Keeling in this discovery.

2.1 *Foundations, Affordances and Moral Perception*

Throughout this thesis, much attention has been paid to the types of agent programs that could yield ethical behavior in artificial moral agents. Often, it seemed that the more these agent programs differed from how human agents apprehend and respond to the demands of morality, the more these programs failed to yield true moral agency, at least as far as the standard view is concerned. To this end, we maintained that in so far as the design goal of current AMAs is to build explicit ethical agents, the necessary division of decisional labor between an AMA and its designer does not make a moral difference, at least in terms of the quality of its moral responsiveness. Indeed, since moral responsiveness—or 'praiseworthy' AMAs—are judged only externally by the human agents affected by an AMA's decisions, it matters little whether the designer or the robot itself decides to act on, say, utilitarian principles, and matters much more whether these principles and the procedures which flow from them expound local normative standards or expectations, or track what morality ostensibly requires in the context. In other words, even an impressive espousal of the mimetic fallacy does not settle the problem of legitimacy endemic to maximalist approaches.

Accordingly, we are left with a comparatively hard definition of a top-down approach to artificial morality, one in which the human programmer decides every step of the type of the claim mitigation strategy he wishes his AMA to expound, from foundations to morally relevant features. This, combined with the idea of the necessarily 'pushless' morality of AMAs and their corresponding lack of options, brings us to our more recent claim that artificial moral agents are maximal moral pull responders.

From a cognitive perspective then, it would seem that maximal moral pull responders, in their perusal of their *Umwelt*, act as ecological creatures, ones whose behavior is directly influenced by the individual claims they encounter. The key to this particular characterization lies in the concept of an *affordance*[46], and its role in what we might call *ecological moral perception*. In his extensive work on animal psychology and perception, James Gibson (to which the term is owed), maintains that "…the affordances of an environment are what it *offers* the animal, what it

[46] Gibson, 2014.

provides or *furnishes*, either for good or for ill…imply[ing] the complementarity of the animal and the environment"[47]. A bit further on, Gibson maintains that the composition, or more precisely perhaps, the phenomenal signature of an object or surface *constitutes* what it affords, implying the rather radical hypothesis that "…the 'values' and 'meanings' of things in the environment can be directly perceived"[48].

Intuitively then, what the concept of affordance implies for the design of the Ethical Valence Theory is that the moral deliberation of a machine can arise *ex datis*—from direct phenomenal experience—rather than *ex principiis*, or the *a priori* evaluation and endorsement of principles of action. In this sense, just as a chair can 'afford' the action of sitting, a helmet-wearing cyclist can *afford* the action of braking, without prior deliberation concerning how, for instance, this particular cyclist's claim to safety tracks utilitarian or Rawlsian principles, how the autonomous vehicle ought to extensionally define these principles given the context, or what sorts of facts underpin this cyclist's claim. Put succinctly, an artificial moral agent, through action in its Umwelt, *perceives* the claims of its environment and acts *directly* in response to them—the focal point of its moral agency is thus its ecological moral perception.

It stands to reason then that the job of the human programmer is to *extensionally define* the moral claims which an AMA will encounter in its *Umwelt,* himself accomplishing the cognitive activity of tracking these claims to relevant moral principles, theories, or expectations. This, in previous chapters, is what we have called the process of artificial moral uptake. In this sense, the division of decisional labor between the programmer and the AMA amounts to the former's complete authority regarding the description and elaboration of the AMA's ecological moral perception. It is also here, however, that a designer can choose to embrace or ignore burgeoning normative convergence, and thus align himself with the principle of total irreproachability.

As we maintained earlier, the arbitrariness of the maximalist's approach to artificial morality flows mainly from his indifference to the acceptability of his chosen claim mitigation strategy in the eyes of individuals, societies or institutions. His chosen moral theory may therefore

[47] Gibson, 2014, 127.
[48] Ibid., p 127.

ignore certain factors or features that people find salient, and in so doing, he will typically revise upon the action-guiding recommendations of common-sense morality. Thus to avoid the charge of arbitrariness, and the problem of legitimacy in the Ethical Valence Theory, the extensional definition of claims must pass through these considerations of acceptability. However, most of the empirical data which is pertinent to claim mitigation strategies looks mainly at the types of morally relevant *features* which could be seen to matter to users or society, or at the very most, the types of morally relevant *factors* which could be pertinent to specific *Umwelts*[49].

In other words, the picture of moral claims which acceptability as moral preference paints is only partially complete—leaving out both the widest and narrowest structural elements of a claim mitigation strategy, or its foundational level and its marks of significance. For instance, the Moral Machine Experiment might have shown that both passenger and societal interest matters, but it did not shed much useful light on *why* these factors matter, or *how* they matter or are instantiated in individuals. Furthermore, the experiment may have established a link between certain morally relevant features and the strength or priority of a claim—for instance, that of a child versus that of a criminal—but these preferences alone did not provide a robust method of claim mitigation which did not collapse into a lamentable form of 'traffic community prioritarianism'[50]. Thus, even if we can lean on the burgeoning normative convergence indicated by acceptability research, it is clear that it does not provide an exhaustive claim mitigation strategy; a statement which seems even more credible if many useful features are obscured by a threshold of moral blindness.

If we then look to moral theory for inspiration concerning how to fill in some of these informational blanks, not just any theory will do. This is because, to respect the principle of total irreproachability, the moral theory from which we take inspiration cannot revise upon common sense morality, a claim which we have made previously. This implies that the theory must be able to accommodate the shape of local common-sense morality which we are using as the basis of

[49] Bonnefon et al., 2016 : Awad et al., 2019.

[50] These missing pieces matter to us here in so far as we are still attempting the design of a top-down, or expert system, which still captures some aspects of normative ethics. If we were not committed to this view, it is likely that a computational social choice-type program would be able to mobilize this information without much further specification (Conitzer, Brill & Freeman, 2015).

normative convergence, and cannot dismiss some of its relevant features and factors. For instance, if empirical research points to the normative relevance of special obligations between a vehicle and its passenger, it is then inappropriate to adopt a claim mitigation strategy which views only the overall goodness of results as normatively relevant. Given that common sense morality is typically seen to underpin a pluralistic account of value[51], we might then expect that very few claim mitigation strategies will hold only one factor to be relevant at the normative level. Indeed, it is likely this particular incompatibility which leads to the common conclusion that the data from acceptability research does not tightly track any pre-established moral theory.

It is precisely this point that leads us into contractualist territory. This is so since, as a school of moral thought, contractualism typically views the correct list of normatively relevant factors as those which would be agreed upon, consented to, or reasonably unobjectionable for suitably disposed and informed individuals acting in society[52]. Further still, it allows some degree of relativism—or 'parametric universalism'—concerning the ways in which particular values are instantiated in different societies, and thus admits contextual variances in responsiveness to claims[53]. The Ethical Valence Theory and contractualism thus share some common convictions. It is in light of this that we could paint the EVT as foundationally contractualist, since what makes it the case that factors and features have moral relevance is precisely this notion of societal agreement and non-objection, rather than say, the overall goodness they bring about[54].

[51] Gert, 2004: Kagan, 1992.
[52] Kagan, 1992: Scanlon, 1998.
[53] "Any plausible moral view would allow for the fact that actions that are right in one place can be wrong in another place, where people have different expectations, or where different conditions obtain. Failing to help a person whose car has broken down, for example, would be a serious wrong in a place where someone who is stranded overnight is likely to freeze to death, but not a serious wrong in a safe country with a mild climate. A view that allows for such variations in what is right, by applying a fixed set of substantive moral principles to varying circumstances, is not relativism but rather what I will call 'parametric universalism'" (Scanlon, 1998, 329).
[54] By contrast, where this similarity is weakest is likely in the substantive concept of 'agreement 'itself. Where contractualism's concept captures a 'hypothetical agreement 'which is seen to form the basis of human thought concerning right and wrong (Scanlon, 1998, 155), our approach here, in relying on empirical data about prevailing moral preferences and attitudes, incorporates more of what Scanlon calls 'actual agreement', which places an accent on prevailing moral consensus and individual acceptance over 'moral correctness', without, as we will see later on, being entirely reducible to it.

In detail, where this similarity is most cogent is likely in the 'dynamic shaping role' that justifiability (or reasonable rejection) accomplishes in individual practical reasoning. As T.M. Scanlon maintains, "There is no fixed list of 'morally relevant considerations' or of reasons that are 'morally excluded'. The aim of justifiability to others moves us to work out a system of justification that meets its demands, and this leads to a continuing process of revising and refining our conception of the reasons that are relevant and those that are morally excluded in certain contexts"[55]. In principle, what this implies for the EVT is that there is no moral preference which is automatically excluded from consideration. In practice of course, it is important that moral preferences are taken to mean just that, rather than external preferences for the moral behavior of others, or preferences which flow from purely prudential reasons.

Additionally, this similarity between contractualism and the EVT does not extend into this 'process of revising and refining', which is certainly carried out by the human programmer, rather than the AMA itself. While the human programmer likely uses rules, principles or specific moral reasons as the evaluative focal point from which he devises moral profiles—in a contractualist spirit—the EVT as an agent program does not simulate this reasoning process. Claims themselves are directly appraised by the AMA, as a condition of its ecological moral perception. This means that the AMA does not deliberate 'across' different principles or profiles, so as to find the one which is least objectionable to the human agents in its *Umwelt.* It does not, in other words, take the diversity consideration to this extreme. Instead, the AMA operates with a single moral profile, which a human agent, e.g. a programmer, user, or passenger, can choose or change.

Finally, this foundational rapprochement between the Ethical Valence Theory and contractualism may be usefully extended to include two deliberative constraints which may help us apprehend some basic structural elements of claim mitigation strategies. Indeed there are two such constraints which are typically seen to exist in contractualist thought, what Derek Parfit has called the *individualist* restriction and the *impersonalist* restriction. The latter pertains to the idea that "…in rejecting some moral principle, we cannot appeal to claims about the impersonal goodness or badness of outcomes"[56], while the former holds that all moral reasons for action

[55] Scanlon, 1998, 157.
[56] Parfit, 2011, 214 : Scanlon, 1998, 222.

"…must appeal to the principle's implications for ourselves and for other single people"[57]. In our terms, these restrictions help to ensure that a) only an *individual*'s complaint can provide the grounds for a claim mitigation strategy's unacceptability, and b) that an argument for the unacceptability of a claim mitigation strategy cannot be made on impersonal grounds, say by citing its failure to maximize everyone's well-being. Together then, these restrictions prevent the aggregation of claims in the AMA's deliberations, ensuring that the AMA only considers direct changes in each individual's degree of claim satisfaction, or degree of moral responsiveness.

2.2 *Claims & Valences*

Despite the Ethical Valence Theory's harkening to contractualism at the foundational level, its structure and deliberative mechanisms are quite exotic to any pre-established moral theory. This exoticism is owed mainly to the basic design motivation of the theory: the will to combine moral requirement and acceptability data in the moral decision-making of an AMA. These two sources of moral content are manifested as *claims* and *valences* in the structure of the EVT, where each individual in an AMA's *Umwelt* is seen to possess both, to a varying degree of strength. In this section, we address each of these concepts in turn.

Analytically, individual claims can be understood as contributory or pro tanto reasons for the AMA's acting in a certain way[58]. Each claim acts as a contributory 'ought', meaning that the strength of a claim is directly relative to how strongly it 'ought' to respond to the individual's claim, or his moral pull. To take an example, in regular driving conditions, an autonomous vehicle has a reason, all things considered, to privilege the claim (to safety) of its passenger in its tactical decision-making. However, when a dilemma situation arises and an unavoidable collision is imminent, the vehicle will be faced with other reasons, all things considered, to privilege the claims of other road users, such as pedestrians or cyclists. Since these reasons are in conflict, the vehicle must then detect which of these reasons is the strongest, and act on the strongest reason. Then, by responding to the strongest claim in its environment, the vehicle is doing what it 'most ought' to do, morally speaking.

[57] Parfit, 2011, 193 : Scanlon, 1998, 229.
[58] Dancy, 2004: Prichard, 2002.

Within the structure of the Ethical Valence Theory then, the role of claims is to capture the contribution that normative ethics could make to AMA decision-making. Claims, in other words, allow the vehicle to ascertain what morality requires in dilemma contexts, by tracking how fluctuations in a morally relevant feature (say welfare or harm) affect the rightness or wrongness of an AMA's action. In many respects, this approach takes inspiration from the 'competing claims' model popular in distributive ethics[59]. However the closest resemblance to our particular concept of a claim is likely found in the work of John Broome, in two respects. Firstly, Broome provides a rather restrictive definition of what can count as a claim, associating it with "…a duty owed to the candidate herself that she should have [some commodity]"[60]. This implies that among all the reasons we could find for awarding an individual some commodity (or benefit or burden), not all reasons will affect the strength of her claim over this commodity. In one sense, this aligns the concept of a claim with the mobilization of purely *constitutive* facts related to a given morally relevant feature[61]. For instance, if 'hunger' is the feature we seek to track, and soup the commodity, then the time elapsed since a claimant's last meal will be constitutive of all soup-claims, but 'most disabled' or 'warmest clothes' will not. In another sense, this ensures that claims respect what is often called the 'separateness of persons'[62]. This means that when deciding which claimant receives soup, the fact that claimant A has an elephant-sized stomach, while claimant B eats like a bird, does not give claimant A a stronger *claim* to the soup, only a reason to decide in A's favor. As Broome explains, "…weighing up [reasons] is the treatment we would naturally give conflicting duties owed to a single person. But conflicting claims are duties owed to different people. Weighing them up, like duties owed to a single person, does not give proper recognition

[59] Nagel, 2012: Voorhoeve, 2014.
[60] Broome, 2017, 195.
[61] In his own work, Broome is frustratingly vague about the distinction between a claim and a reason. His best answer, which we see as aligning with our concept of constitutive facts given its reference to Nozickian entitlement theory, goes as follows: "Suppose we are interested in the distribution of income between people. For each person, there are reasons why she should have some income…If the person would derive some benefit from income, that is a reason why she should have some. If she has earned some income, that is another reason. Some of these reasons may be claims, and others not. There is scope for a great deal of disagreement about this. One view is that claims can only arise historically, through the process of trading and contracting. Another is that everyone has an equal claim to good. And there are many other possible views" (2017, 197).
[62] Rawls, 2009, 22-27.

to the people's separateness"[63]. Broome's notion of a claim is then complementary to Scanlon's individualist restriction mentioned in the previous section, and thus to the EVT.

Secondly, Broome places significant emphasis on the principle of proportionality in managing claim conflict:

> When claims conflict, I suggest that what fairness requires is not that they be weighed against each other and other reasons, but that they actually be *satisfied in proportion to their strength*…Stronger claims require more satisfaction and equal claims require equal satisfaction. Also, weaker claims cannot simply be overridden by stronger ones: if a stronger claim is satisfied to some extent, then so should a weaker one be to a lesser extent[64].

While Broome uses divisible goods throughout many of his examples of claim conflict, and thus departs from the material conditions under which many AMA's will operate, there are still a number of useful associations to be made here. Firstly, since Broome's preoccupation with claims is born mainly of a will to advocate for the fairness of lotteries, his concept of equal claims aligns with our idea of a threshold of moral blindness, or the point at which it is acceptable to randomize responsiveness across claimants. Indeed, Broome counts having a fair stake in a lottery as a type of 'surrogate' claim satisfaction[65]. Thus, while we arrive at the use of lotteries via very different justifications, our concept of equal claims and their management aligns with Broome's. Secondly, the idea of moral responsiveness as a question of *proportionality* to claim strength is one that operates at the heart of the Ethical Valence Theory. In one sense, this concept aligns with the Scanlonian *impersonalist* restriction, since we are not concerned with the absolute level of claim satisfaction. This means that instead of seeking to minimize total hunger in a decision context, we are instead concerned with how the awarding of soup affects the hunger of each claimant, considered individually. In another sense, this type of proportionality prevents a more *exclusivist*

[63] Broome, 2017, 195.
[64] Broome, 2017, 196.
[65] "Unfairness is almost inevitable when there is not enough of an indivisible commodity to go round everyone who has a claim. But I believe a lottery can mitigate that unfairness. People cannot all get the commodity in proportion to their claims, but they can at least have a chance at getting it in proportion to their claims. Having a chance, I believe, is a sort of surrogate satisfaction of the claim. This explains the fairness of a lottery." (Broome, 2017, 196).

satisfaction of some moral claims, such as what might happen when an autonomous vehicle runs over a pedestrian to avoid giving its passenger whiplash. What this implies is that no claimant is given absolute priority in the decision-making of the AMA, regardless of the strength of his claim.

Succinctly then, a claim within the Ethical Valence Theory constitutes (i) a moral entitlement owed to an individual, (ii) that is indicative of some morally relevant feature (e.g. harm, hunger), (iii) that is underpinned by constitutive facts (e.g. physical integrity, time since last meal), and (iv) that increases in strength in light of its relation to other claims (e.g. most injured, hungriest). The goal of the EVT is then to distribute its moral responsiveness in proportion to each claim, and then to the strongest claim when not all can be satisfied.

With the idea of a claim firmly in place then, we may now address the concept of a *valence*. If moral claims provide a point of entry for strict responsiveness to important moral features such as harm or welfare, the notion of a valence is meant to capture the space left over for the minimalist's ideal. In this sense, if a claim is underpinned by constitutive facts, entitlements and moral duties, a valence is underpinned by scalar facts, and moral reasons and features which while pertinent, nevertheless fail to be constitutive of a moral claim. Most generally then, the *raison d'être* of a valence is to maximally capture our considered moral judgements concerning AMA behavior in an *Umwelt*, and in this way allows the EVT to respect the shape of local common-sense morality in its moral responsiveness.

Until this point, we have entertained the notion of a scalar fact primarily as a function of the types of moral preference that flow from empirical studies such as the Moral Machine Experiment. For example, it is a scalar fact that people generally prefer to sacrifice criminals over pregnant women in unavoidable collision scenarios. In theory however, scalar facts and thus valances can accommodate a much broader set of pro tanto moral reasons for action: tracking an original equipment manufacturer's preferences for liability and responsibility, the personal preferences of a passenger, industry norms concerning the treatment of vulnerable road users, or various institutional recommendations concerning ethics settings. In effect, the exhaustivity of valences is limited by only three parameters.

First, and similar to moral claims, valences can be severely capped by the threshold of moral blindness which flows from the further feature problem and perverse incentives. In this sense, while they are conceptually possible to track, many scalar facts relating to elements such as socio-economic status, race, or criminal background will not be acceptable options to underpin valences[66]. Secondly and relatedly, depending on the epistemic resources available, valences might be further capped by the phenomenal limitations of an AMA. What this means is simply that an AMA must be able to detect whichever features underpin the valences of its environment, which in embodied AMAs at least, will then be a function of its object detection and classification algorithms, or perhaps its connection with the Internet of Things. In this sense, while it might seem useful to know whether a soup-claimant has a five euro bill in his pocket, the availability of this information is very much dependent on the ontological conditions of the AMA, or its ability to see through clothes, read the claimant's bank statements, or ask pointed questions.

Finally and most importantly, the preferential data afforded by the scalar facts under consideration must admit of some ranking, hierarchy, or order of priority. This is so since valences, just like claims, vary in strength and are particular to each individual; this time in relation to that individual's alignment with the scalar facts under consideration. Thus, taken by themselves, valences will always admit of a form of prioritarianism, organizing an *Umwelt*'s bearers of moral value into valence categories of relative priority. The more an individual reflects these facts, the stronger his valence becomes. This need for a hierarchy flows less from a desire that the EVT spare the darlings of society at all costs, but rather, that in situations of conflict across *equal* claims, a scalar fact may be able to break the tie, rather than a purely randomized decision. In this sense, if both a passenger and a 12-year-old pedestrian are equally likely to be injured as a result of an autonomous vehicle collision, if scalar facts support a stronger valence for children, then the

[66] This is indeed precisely the type of distinction which contractualism aims to capture in its concept of reasonable rejection: "My reason for rejecting a principle might be, not so much that it imposes a certain burden on me, but the way in which it imposes that burden—and what that principle thus says about me. For instance, consider a principle that allocates benefits and burdens on the basis of race, and contrast this with a principle that allocates the same benefits and burdens randomly. I cannot reject the racist principle simply because of the burden it imposes on me—after all, the random principle imposes an identical burden on someone else. Rather, I reject the racist principle because, by regarding my race as a relevant ground for the distribution of benefits, it imposes a burden in a way that constitutes failure to respect my status as a person" (Ashford & Mulgan, 2018).

vehicle will choose to spare her over the passenger[67]. In some ways then, valences which track preferences such as *vulnerability* into broader categories—i.e. bicycle, pedestrian, truck or other road user types—might not only sneak past the threshold of moral blindness, but might additionally have a positive impact on an AMA's public perception as a praiseworthy agent.

The fact that much research into moral preference and acceptability has focused on highly contentious, and somewhat 'bias-pinpointing' criteria should thus not deter machine ethicists from the larger project of ensuring that robots behave in ways we expect them to in morally salient contexts. While it certainly feels morally questionable whether the wealth or criminal background of a person should make a difference in a decision which might end their life; the fact that the person is a child or is more vulnerable to lethal injury may indeed matter morally, in ways which are captured by societal values, and are incorporated with great difficulty into many pre-existing moral theories. The root of the problem is not so much that it is wrong to render robots sensitive to these attitudes, or that in doing so, they will fail to maximize the good they can produce, or the evil they can avoid. Instead, the problem flows from the simple truth that robots do not have their own values and principles, and thus it is the role of society—with all of its conflicting ends, values, and conceptions of the good—to decide how these agents should act in contexts where there are no resoundingly acceptable answers.

The 'valence' of an individual then, is not a pure representation of his worth to society, but rather a representation of how his particular material and contextual circumstances affect the vast web of stakeholder interests that underpin the implementation of a given AMA—including not only users, but also original equipment manufacturers, regulatory bodies, and international institutions. In other words, within the Ethical Valence Theory, valences provide an opportunity for the decidedly grassroots and democratic appeal of bottom-up sources of moral content to *influence* the moral claim of an individual, without *supplanting* this moral claim.

[67] It is important to specify that beyond situations of conflicting equal claims, a valence can never trump a claim, leading to a situation where societal moral preference has the final say in dilemmatic decision-making. As was explained previously, the EVT instructs an AMA to respond to the strongest *moral claim* in its environment, when such a claim exists. We will explore this in greater detail in the next chapter.

Thus, while the EVT cannot completely escape the contentious use of the 'moral wisdom of the crowd', it is important to recall that this wisdom *alone* does not decide how the AMA acts towards human agents. Indeed, if acceptability research can be conducted along less contentious lines, the use of scalar facts in artificial morality design will likely lose much of its bite. Conclusively then, it is important to recall that it is up to the human designer to decide which facts and features enter into the constitution of an individual's valence, and also whether and how these valences are used in claim mitigation.

2.3 *Moral Profiles*

The final conceptual piece of the Ethical Valence Theory is the notion of a 'moral profile': a specific decision procedure or method which mitigates the different claims and valences of individuals. Essentially, each moral profile provides a specific criterion of rightness: a maxim or rule which decides the rightness or wrongness of action options. In this way, a moral profile also dictates which claims the AMA is sensitive to and when, and how those claims are affected by an individual's valence strength. Of course, the types of moral profiles which constitute desirable claim mitigation strategies will vary wildly with the type of *Umwelt* in which the AMA acts. This is because the trio of claim-valence-moral profile is meant to circumscribe the ideal pattern of moral responsiveness in a given context, and bluntly, different contexts often demand very different forms of responsiveness. Here then, we will only address how the flexibility of such an approach can help to quell some of the persistent problems in the design of artificial morality, while leaving a more technical discussion of moral profiles to the next chapter.

To begin, it might be useful to return to the concept of surrogacy in artificial moral agents. In previous chapters, we have seen how many different forms of AMA technology can be seen to have a user-centric purpose-oriented ontology, which is to say that at least part of their role as a practical agent is to act on behalf of, or in the interest of, a particular individual. Given our later analysis of the maximalist position, we may notice how the 'optionless' application of standard moral theories conflicts with this purpose, yielding agents who give away their user's groceries, murder their aunts, or donate their money to charity without consent. What underlies these rather extreme examples is a structural problem in the reasoning of these agents: their moral theoretic

vision of moral responsiveness requires them to act as a *distributive agent*, one who equally considers all moral claims, and who has no special obligations to particular individuals. This, as we have maintained, poses a serious problem for the *adoptability* of these agents, since it is at least plausible that many machine decisions which neglect their user's interest will be perceived as instances of machine misconduct by this user. One way to address this problem is by augmenting the strength of the principal user's valence, thereby simulating a form of morally admirable partiality in the AMA, and ensuring his priority in situations of equal claim conflict.

Another, more structural route however, consists in making categorial separations in the types of moral profiles destined for surrogate agents. Rather vaguely put, this involves delineating different spaces or levels of moral concern across the bearers of moral value with which the machine interacts, painting claim mitigation as the resolution of conflict between 'sets' of claims. A very clear example of this can be seen in autonomous vehicles, where the claims of the passenger(s) (those individuals inside the vehicle) are balanced against the claims of those individuals external to the vehicle. More imaginatively perhaps, Leben's personal shopper robot might also benefit from such a categorial separation, since the claims of its principal user to her groceries are likely consistently at odds with the claims of other individuals to sustenance.

The principal benefit of these sorts of categorial separations is the possibility of a *threshold-type* approach to moral reasoning. This implies that, rather than uniformly satisfying one moral claim at the expense of another, the AMA is able to satisfy one individual's claim *up to a point* or under certain conditions, and another set of claims when conditions change. In some ways, this approach formally resembles what Derek Parfit has called 'sufficient altruism'[68], or what is known in population ethics as 'critical level utilitarianism'[69]. Intuitively, sufficient altruism tracks the interplay between prudential and moral reasoning, where a rational self-interested agent can be seen to respond to *some* but not *all* of morality's demands, thereby restricting him from acting as a pure altruistic agent; a concept which mirrors that of the problem of demandingness in moral

[68] "By 'sufficient altruism' I mean sufficient concern for others, where the limiting case is impartial benevolence: an equal concern for everyone, including oneself." (Parfit, 1984, 66).
[69] Blackorby, Bossert & Donaldson, 1997: 2005: Broome, 2004. Parfit also addresses this concept in what he calls the *appeal to the valueless level* (Parfit, 1984, 412).

theory. The latter theory, in turn, points to the idea of a critical threshold of individual well-being below which life is not worth living, regardless of how many more lives are lived at that level[70].

What these categorial distinctions give us, in other words, is a deliberative method for artificial morality which from an external perspective at least, likely resembles a *compromise* between the expectations of human agents. In this sense, a personal shopper robot could be endowed with a 'critical threshold of grocery guardianship', thereby ensuring that its user receives a portion of her groceries, but that any remaining goods may be distributed to others, for instance in accordance with the maximin principle. Put another way, once the special obligations an AMA has to its user are satisfied, the machine is free to act as a purely distributive agent in its environment, according to any number of potential criterions of rightness. The user is assured a certain share of the AMA's moral responsiveness, beyond which the AMA is free to act on 'optionless' moral principles in its satisfaction of moral claims.

It is thus within the structure of a moral profile that the *practical push* of an AMA is likely best accommodated, and importantly, this accommodation does not prevent the AMA from responding to what morality requires. Moreover, in so far as the abstract notion of a 'critical threshold of moral responsiveness' is not itself tethered to any particular conception of the good or the right, it provides an interesting opportunity for user input or meaningful human control. Thus, this threshold could be represented as an 'ethical dial' that mitigates how responsive a user's AMA should be to the moral claims of others, thereby avoiding the more paternalistic option of mandatory ethics settings in such machines.

Here again, we might be tempted to perform a maximalist calculus concerning the collective effects of machines which accomplish such compromises, and thereby fail to maximize the potential moral benefits of their actions. Indeed, when assessing AMA behavior from the perspective of moral philosophy, it appears that we are failing to capitalize on an opportunity to

[70] By way of analogy, we could say that in a milk-maximizing world, it is not better that 50 000 bottles with only a drop of milk should exist, rather than 500 nearly full bottles, even if the 50 000 bottles together contain more milk. Establishing a critical threshold for milk volume here allows us to avoid the 'repugnant conclusion' that 50 000 bottles are better, one which follows from aiming only at the pure maximization of milk quantity in the world.

improve the moral efficiency of society, an end which seems rather indisputably good. However it should be equally clear that from a pragmatic and more humanistic view, machines which exhibit sensitivity to the cultural and contextual nuances of human beings, and behave in ways which respect their *moral value* rather than their *value to morality*, are likely praiseworthy in exactly the sense that artificial morality was originally meant to capture. Then, even if there are morally inferior human villains in the world, it does not follow that it is technology's job to correct this fact, or *a forteriori*, that pure robotic adherence to a given moral paradigm will get this job done.

3. *Conclusion*

In this chapter, we have attempted to question whether the original intentions behind the implementation of moral behavior in artificial moral agents truly translates into the maximalist project of pure adherence to moral theory. Our main bearing in this pursuit has been the concept of *total irreproachability*, where irreproachability is not best understood as freedom from physical harm, nor the assurance of the 'best' or 'correct' events and outcomes judged from a view from nowhere. Instead, what we have attempted to uncover is a concept of harmfulness that tracks non-adherence to the types of norms that human agents already respond to, or in cases where these norms are less clear, the types of expectations they have empirically been seen to hold. In this sense, the job of artificial morality was not that of responding to the 'right' moral requirement in conflicting cases, nor was it that of transforming social contexts into opportunities for human moral improvement via technological means. It would seem instead that a truly praiseworthy AMA requires something more akin to a courteous and considerate disposition, displaying a sensitivity to the ways in which human agents live and work, and the ways in which they wish to be treated.

Our conceptual elaboration of the Ethical Valence Theory, in turn, can be seen as our answer to this call for praiseworthy AMAs. One of the principal intuitions behind this theory is that total irreproachability, or praiseworthiness likely sits somewhere in-between common-sense morality and pure moral requirement, or somewhere between what is acceptable and what is morally right. Another intuition is that the very existence of these machines is itself contingent on a host of considerations and constraints which have little to do with morality proper, and much more to do with the material realities of the technology sector and society at large. In this sense,

just as artificial moral agents might do well to compromise in their responsiveness to claims, so too might machine ethicists do well to compromise on their unflinching espousal of moral theory. Thus, through the atomic skeleton of a claim, valence and a moral profile, new and contextually sensitive forms of moral behavior can take form, ones which are not suited for human moral agents with value and a moral push, but rather are suited for entities which respect these types of agents in conflictual situations. In other words, the Ethical Valence Theory paints artificial morality as an exercise in *complementarity*—rather than *mimicry*—with the human condition. The next and final chapter of this thesis explores how this approach might take form in a particular type of AMA, autonomous vehicles.

CHAPTER

7

The Ethical Valence Theory & Autonomous Vehicles

Throughout this thesis, we have relied on many rather fictitious examples of artificial moral agents to flesh out various moral intuitions about their purpose and behavior—meatpacking robots, berrypickers, soupsaints and Aunt Agatha Terminators to name a few. In one sense, this imaginative treatment of AMAs and their implications flows from a methodological reliance on what Daniel Dennett has called 'intuition pumps', cleverly constructed artificial cases which enable us to isolate whichever factors or features are salient to our analysis, a practice which lends itself well to philosophical reflection. In another sense, however, our use of imagined cases reflects a more empirical truth about the subject matter of machine ethics: namely, that the design and elaboration of artificial morality remains something of a *prospective* venture, addressing machines which are not yet on the public market, or which have only just begun to be ubiquitous members of various human social contexts. In other words, it is often the rise and *trend* of automation that leads us to a concern for the ethical behavior of these machines, rather than a plethora of real-life examples in which a lack of moral behavior has led to disastrous consequences.

This prospective account of machine ethics is challenged however, by one particular type of AMA, the autonomous vehicle (AV). It would seem that the autonomous vehicle—at least when

understood as an engineering ideal—has tickled the collective imagination for some time indeed. Arguably, the autonomous vehicle was born in the 16th century, when Leonardo Da Vinci first conjured the design of a self-propelled cart that was able to navigate a pre-determined course without human intervention. This same ideal cropped up quite a bit later in the design of the Stanford Cart in 1961, an autonomous vehicle designed to navigate the moon's surface by following a drawn trajectory, thus obviating the need for complicated tele-communication technology. However, the birth of the modern autonomous vehicle likely dates back to the DARPA challenges of 2004-2013, where the United States government invited various robotics companies to race their best AV prototypes; first across a 150 mile stretch of desert highway, and then in successive years, across various types of urban and dynamic environments[1]. Today, autonomous vehicles are on the edge of ubiquity, and can be seen cruising the streets of Palo Alto, California, or operating behind the 'autopilot' features of many late model cars, and are projected to garner a significant place in mainstream commercial markets in as early as 2030[2].

Correspondingly, autonomous vehicles are familiar to both public and private spheres in ways that many other types of AMA are not. Still, this rather basic understanding of autonomous vehicles eludes a number of important facts. First, it should be noted that today's AVs are subject to an *incremental* form of implementation, a notion which is captured by the so-called 'levels of automation' that serve as industry standards for their development[3]. Indeed, of the five levels available, only the 5th constitutes a *truly* autonomous vehicle, one which can operate in any type of environment or decision context, and never delegates control back to the passenger. Other levels (1-4), require some degree of on-board supervision in challenging conditions such as storms, heavy urban traffic, or imminent collisions[4]. Unsurprisingly perhaps, the intellectual interest of machine

[1] Buehler et al., 2009.
[2] Litman, 2017 : 2020.
[3] SAE, 2016.
[4] In detail, the five levels are laid out as follows: Level 0 vehicles constitute traditional cars, where a human driver performs all driving tasks. Level 1 have at most one assistive feature, such as cruise control. At level 2, there is a more significant degree of automation, where multiple assistive features can run concurrently. Level 3 autonomous vehicles are the first to have a genuine automated driving mode, but this mode can only operate under ideal conditions. Level 4 vehicles have high automation, where most of the driving tasks are accomplished by the vehicle, and where human supervision is not required. Still, the vehicle can delegate the driving task back to the passenger if unmanageable conditions arise. Finally, level 5 constitutes full automation, where all driving tasks are delegated to the vehicle in all conditions, and no supervision or intervention is required on the part of the passenger.

ethicists has concentrated mainly on this fifth level of autonomy. One likely reason for this is that the vehicle is delegated tactical control in collision scenarios only in level 5 automation, and thus it is only here that a vehicle might make decisions which pose a serious risk of harm[5].

Secondly, while it is projected that AVs will be available for commercial purchase relatively soon, wide-spread adoption, and thus the emergence of an AV-*dominated* traffic environment, is not expected to occur until much later[6]. In this sense, there exists a second, *temporal* type of incrementalism in autonomous vehicles, one relating to the shift from predominantly human drivers, to a mixed fleet, to the eventual elimination of human-driven vehicles from the world's roads. It follows from this that the *Umwelt* of an AV can vary significantly in function of its location on this incremental spectrum. Initially for instance, autonomous vehicles must be equipped to interact in a mixed fleet, where many human drivers and passengers are unaccustomed to their presence and behavior. Much later however, it may even be considered immoral, or certainly in bad taste, for a human agent to drive his own vehicle, since his driving acumen will likely be far outstripped by highly advanced and communicative AVs[7]. For our purposes at least, the principal upshot of this temporal incrementalism is the idea that many AV design choices will tend to be temporary rather than absolute. For example, forms of interactivity that seem appropriate for initial phases of implementation—such as the use of specific lights, sounds and even expressions to indicate the vehicle's behavior to surrounding pedestrians—may seem archaic once the technology is more widely adopted.

A more interesting claim is made, however, in asserting that permutations across both incremental scales may affect what *morality* requires in AV behavior[8]. This seems plausible if we

[5] Another reason is that due to the inherent dangers of delegating driving tasks in critical situations such as inevitable collisions, i.e. the passenger's lack of awareness or preparedness to take over control (Walch et al., 2015: Casner et al., 2016), the expected end-goal of AV implementation is likely full automation (Sparrow & Howard, 2017).

[6] For instance, true market saturation is not expected to occur before the 2070s (Litman, 2020).

[7] "If vehicles without a human being at the controls are safer than vehicles with a human being at the controls, then the moment a human being takes the wheel they will place the lives of third parties—as well as their own lives—at risk. Moreover, imposing this extra risk on third parties will be unethical: the human driver will be the moral equivalent of a drunk robot. Eventually, we believe, the compelling moral argument against human drivers will be reflected in law: driving will be made illegal" (Sparrow & Howard, 2017, 212).

[8] This topic is given extensive treatment in (Nyholm & Smids, 2018).

can accept that normative convergence surrounding ideal AV behavior can strengthen as AVs saturate the traffic community. This implies not only that road users themselves will have a better idea of what to expect of an AV once they are ubiquitous, but also, that legal and scientific convergence will already have provided answers as to the attributability of AV decisions, or who is liable or responsible for them. In this sense then, many things that could be seen to be morally relevant to AV actions—who is responsible or to blame for the accident, who made the decision as to its programming, or what types of road users are typically involved in collisions—may vary significantly across time. Thus, it behooves one to be temporally precise when making moral claims about autonomous vehicles. To this end, we will focus on the early implementation phases of level 5 autonomous vehicles in our analysis here, a time when human drivers are predominant, and where normative convergence is burgeoning rather than clearly defined.

Finally, and perhaps most importantly, the persistence of human drivers in traffic environments impacts what is likely the principal expectation that drives AV development: a reduction in the number of road-traffic deaths[9]. In effect, the number of deaths on the world's roads remains unacceptably high, with an estimated 1.35 million people dying each year and up to 50 million injuries[10]. According to the most optimistic of available predictions, autonomous vehicles are projected to reduce this number by a startling 90%[11]. The key to this prediction lies in the autonomous vehicle's behavioral optimality, and subsequent potential for removing human-error-related accidents from the world's roads[12]. In other words, since autonomous vehicles cannot drink and drive, text and drive, or drive aggressively or distractedly, many of the root causes of traffic fatalities will be eliminated. Add to this additional indirect benefits such as a reduction in pollution and traffic congestion[13], increased accessibility for the handicapped or elderly, or the ability to engage in on-board activities during commuting[14], and autonomous vehicles almost appear to be an exercise in technological philanthropy.

[9] "A major reason why there is so much interest in autonomous vehicle development by car manufacturers—and so much support for this development from governments worldwide—is the societal benefits that autonomous vehicles are projected to have…This has led one commenter to suggest that self-driving cars will save more lives than world peace" (Gurney, 2015, 191).
[10] World Health Organization, 2018.
[11] Airbib & Seba, 2017: Fagnant & Kockelman, 2015: Gao, Kass, Mohr & Wee, 2016.
[12] Fleetwood, 2017, 532: Gurney, 2015: Goodall, 2014: Beiker, 2012, 52.
[13] Fagnant & Kockelman, 2015.
[14] Anderson et al., 2014.

However, while the long-term impact of AV implementation may indeed be that of rendering the world's roads virtually 'accident-free', it would be a mistake to consider autonomous vehicles as purely innocuous road users. While it may be the case that a *pure* AV fleet may avoid many if not all potential accidents—firstly in virtue of their superior observation and reaction to the traffic environment[15], and secondly in virtue of the types of vehicle-to-vehicle communication (V2V) or vehicle-to-device (V2D) communication with which they are equipped[16]—the more pressing realities of *mixed* fleet traffic require that autonomous vehicles interact with human road users who may either be unaccustomed or unable to interpret the behavior of an AV, or communicate effectively with that AV[17]. In a different sense, it also seems plausible that while AV's may avoid *human* errors in their driving, they may not themselves be completely error free[18], a concern which is likely borne out by cases like that of the object classification error which lead to the death of Elaine Herzburg in 2018.

Generally then, it would seem that the modern autonomous vehicle differs from its predecessors in ways which appear to grant it entry into the constituency of artificial moral agents. Mainly, this is due to the fact that a) autonomous vehicles act in contexts where human agents are present, but also b) that the interests and welfare of these human agents can be directly affected by the actions of the vehicle, and c) that these interests are in conflict. Indeed, this conflictual structure of AV decision-contexts frustrates the rather straightforward goal of harm reduction in AV decision-making, since it is likely that given sufficient decisional autonomy, autonomous vehicles will need to make *sacrificial* decisions which result in significant harm, if not death, to one or more

[15] "Unlike the human driver, the autonomous vehicle 'sees' everything, in the vicinity; reacts at speeds humans cannot match; and constantly checks the performance of every component in the vehicle to ensure that it is functioning properly" (Gurney, 2016, 191).
[16] This technology allows 'connected' vehicles to share information such as real-time traffic data with other networked AVs (Glancy, 2012). It might also permit a significant amount of data sharing concerning the facts and features of human road users, if privacy is not adequately addressed in AV development.
[17] To be sure, many original equipment manufacturers are cautious about these risks, a disposition which leads them to make rather creative choices when it comes to designing for road-user safety. By far the most comical of these designs is Google's patent for car hoods "…sticky enough to lock onto a pedestrian the car hits, the assumption being that sticking to the front of the car beats getting blasted back into an intersection" (Eifling, 2016). Presumably, the vehicle could then 'transport' this startled crash victim straight to the hospital.
[18] Goodall, 2017.

road users[19]. These realities, combined with the unique opportunity these vehicles afford to perform a deliberative *decision* in such cases—rather than an unreflective reaction as would be the case in human drivers[20]—has brought many authors to establish a link between AV decision-making and the so-called 'trolley problem'[21]; the structure of which aligns with what Geoff Keeling has called the 'moral design problem'[22], as per our discussion of narrow artificial morality in chapter IV.

This harkening of autonomous vehicles to run-away trolleys has been met with a fair amount of criticism in the literature however, convalescing mainly around the inadequacy of this model to capture what must be accomplished in a truly robust programming of moral behavior in AVs. The point of departure for these authors is typically skepticism surrounding the claim that these types of trolley cases will actually arise in real-life AV decision-making[23], and thus that we should consider answers to the trolley problem *because* they shall arise[24]. Two popular objections to this claim revolve around what Geoff Keeling has called the 'Not Going to Happen Argument', and the 'Moral Difference Argument'[25].

The former argument relies on the impossibility or rarity of dilemma-type cases in real-life collisions to cast doubt on the cogency of the trolley problem. As Noah Goodall maintains, "Engineers working on vehicle automation are often asked about the trolley problem. The most common response seems to be that trolley problems are avoidable, implausible, rare, and distractions from more productive efforts. They are considered avoidable because in many trolley problems, the vehicle must decide how best to crash when, with the right sensors and algorithms, the situation should have been avoided entirely"[26]. This forces the AV ethicist to argue for the

[19] "The main safety goal for any driver—human or machine—is to avoid harm. Unfortunately, both humans and today's best computers are imperfect at it." (de Freitas, Anthony & Alvarez, 2019, 4).
[20] Lin, 2016.
[21] Thomson, 1985: Foot, 1967.
[22] Keeling, 2017: 2018: 2020a: 2020b.
[23] Keeling, 2020a: Himmelreich, 2018: Nyholm & Smids, 2016: Nyholm, 2018a: 2018b.
[24] Lin, 2016: Leben, 2017: Goodall, 2014: 2019.
[25] Keeling, 2020a: 2020b.
[26] Goodall, 2019, 3. See also (Roy, 2016). Johannes Himmelreich, in a slightly different vein, argues that it is impossible for an AV to encounter a trolley-type collision while still maintaining a level of control required to make a deliberative decision about how to crash (2018, 673-4).

focus on trolley cases on rather prudential or precautionary grounds, a move which is often made stronger by the elucidation of more 'realistic' cases which still retain the dilemmatic nature of the original problem[27].

Even if the cogency of trolley problems is somewhat questionable from an empirical perspective, this is likely not the only motivation that drives engineers to avoid its use and promotion in connection with autonomous vehicles. Indeed, the 'Not Going to Happen' argument itself seems to be substantiated by the expected behavioral optimality of autonomous vehicles, in so far as superhuman AV drivers should likely be able to avoid accidents such as these. Admitting that these vehicles could crash might then hamper public adoption or acceptance of the technology. But at a deeper level, it seems plausible that this unwillingness to engage with the potentiality of driverless accidents flows from an unwillingness to pronounce any positive, deliberative solution to such cases, especially in the public media. This reticence is likely the consequence of a particularly rough encounter of this kind, one that occurred when Mercedes executive Christophe von Hugo fatefully claimed "…Save the one in the car…If all you know for sure is that one death can be prevented, then that's your first priority"[28] in an interview with Car & Driver in 2017.

In a matter of days, the media backlash for these 'killer cars' became so virulent as to force von Hugo to reverse his position down to quasi-neutrality, by reissuing a statement that disallowed the legal possibility of weighing human lives in the first place[29]. It seems fair that the humility of his initial premise—that an AV should protect the lives of those over which it has the most direct control, which in von Hugo's mind, was the passenger of the AV—was completely lost in the resulting media debate, giving rise to Mercedes' damaging reputation as a car company willing to kill the poor in order to protect the wealthy. It is then far from shocking that major original equipment manufacturers have since actively evaded such public discussions and declarations. Unfortunately, in the years following, this tight-lipped attitude has damaged any prospective

[27] Goodall, 2014: 2019: Keeling, 2020a: Lin, 2016.
[28] Taylor, 2017.
[29] In effect, he endorsed the 'not going to happen argument' to near perfection: "This moral question of whom to save: 99 percent of our engineering work is to prevent those situations from happening at all. We are working so our cars don't drive into situations where that could happen and [will] drive away from potential situations where those decisions have to be made" (Taylor, 2017).

normative convergence around the ideal ethics policy for AVs, and if anything, has galvanized the association between trolley cases and AVs in the public's imagination. Thus, it would seem that even if the empirical realities of AV collisions look nothing like trolley cases, car manufacturers may still need to publicly solve this problem in order to garner acceptability for their products.

On an entirely different tangent, the use of the trolley problem has come under criticism for what Geoff Keeling has called the 'Moral Difference Argument'. This latter argument concerns the idea that the trolley problem, construed as an abstract thought experiment, is blind to many features and factors that human agents, and perhaps programmers and policy makers will likely find to be morally relevant in AV decision-making[30]. These features can include aspects such as moral blame or responsibility, but also special obligations or ties that an AV may have in regard to certain road users. In a slightly different vein, some authors have held that the types of moral intuitions we would have in regards to the trolley problem differ in salient ways from the moral intuitions made in real-life dilemma contexts; a claim most often substantiated by pointing out the differences that risk and uncertainty can make in judgements about what morality requires[31].

In response, we might align with Keeling in maintaining the difference between the use of the trolley problem as a deliberative *model* for AV decision-making, rather than a tool by which our moral intuitions about AV programming can be laid bare. In this sense, if our aim is to illicit the types of 'axiological commitments' AVs should likely exhibit in their decision-making, the fact that a perfect practical correspondence to these cases is rare, or that these cases deal in certain outcomes rather than probabilities, does not seem to significantly impact our search for the types of morally relevant factors and features which could inform what morality requires in AVs[32]. In

[30] Keeling, 2020a: Nyholm & Smids, 2016, 1282-4.

[31] Himmelreich, 2018. As Nyholm & Smids maintain, "Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant. And what we pick out using these concepts are things within different metaphysical categories, with different modal status…" (2016, 1286). Jan Gogoll and Julian Muller (2017) defend a compatible view which finds the disinterested perspective inherent to the trolley problem to be fundamentally different from the more personal, agent-oriented perspective of those affected by AV accidents, and for this reason doubt its cogency for AV ethics setting design.

[32] "What I am claiming is that claims about the morality of risky prospects are nonsensical in the absence of an underlying axiological commitment that *something* matters; and considering hypothetical non-risky cases may inform our axiological commitments" (Keeling, 2020b, 62).

other words, the fact that harm to human agents counts as a morally relevant feature of AV decision-making does not seem to be significantly disturbed by the fact that AVs may only rarely cause harm, or that this harm is probable rather than certain. The problem lies not with the trolley problem as an intuition pump, but rather with what programmers decide to do as a result of engaging with this intuition pump, and how this in turn comes to affect the way artificial morality is designed in autonomous vehicles.

To this end, we have addressed many such responses to the trolley problem, and to the ethical problems of autonomous vehicles across previous chapters: Leben's Rawlsian algorithm, MIT's Moral Machine Experiment, the concept of mandatory or customizable ethics settings, traffic community prioritarianism, the moral design problem, among others. In this sense, it seems superfluous to rehearse these ideas anew in this chapter, especially since they constitute some of the major—or at least most robust—approaches to AV ethics available in the literature. Instead, we will provide a brief perusal of some of the main themes here, and address more precise points as they crop up in the computational presentation of the Ethical Valence Theory. In this way, we will analyze the deeper intuitions of these authors, rather than the models and theories for which they advocate.

Perhaps the best point of departure consists in addressing the main point of consensus among AV ethicists: that the notion of *human harm* is the principal morally relevant feature of AV collisions[33]. Indeed, this is likely what brings Geoff Keeling to view the moral design problem of autonomous vehicles itself as an exercise in harm allocation[34], and what gives an AV's allegoric connection to the trolley problem so much bite. Further, there appears to be relative consensus surrounding what constitutes human harm: damage to the physical integrity of a human being; the probability of survival of a given individual[35], and the estimated severity of injury in a given

[33] Keeling, 2017: 2020b: de Sio, 2017: Goodall, 2014: 2016: 2019: Lin, 2016: Evans et al., 2020: Gogoll & Mueller, 2017: Leben, 2017: Gurney, 2015: Bonnefon et al., 2016: Awad et al., 2019: Contissa et al., 2017: Hubner & White, 2018: Himmelreich, 2018: Nyholm & Smids, 2016: Nyholm, 2018a: 2018b: Sparrow & Howard, 2017: Gerdes & Thornton, 2015.
[34] Keeling, 2017: 2018: 2020a.
[35] Leben, 2017.

individual[36] being two popular marks of significance. Proposals in the literature can then be usefully divided according to how authors propose to theorize and eventually allocate this harm.

To this end, significant attention has been spent on the idea of general harm minimization (GHM) as the correct criterion of rightness for AV collision algorithms[37]. This has less to do with a theoretical commitment to consequentialist or utilitarian moral theory, and more to do with a rather obvious alignment with the purported ends of AV technology: an increase in road user safety, or a reduction in traffic fatalities. In this sense, many AV ethicists use GHM as something of a tacit end, or as an analytical ground-floor upon which they build their specific approaches.

In this vein, Jan Gogoll & Julian Muller take GHM to be the tacit end of AV ethics settings since they view this to be the socially optimal outcome of AV ethics settings, and use the GHM to justify the imposition of mandatory ethics settings (conducive with GHM) following a contractarian justification[38]. Similarly, Jeff Gurney, Patrick Lin, and Bonnefon and colleagues all assume that "society will want the autonomous vehicle to minimize the amount of harm that results from an accident, regardless of who is at fault"[39], using this to entertain naive views of utilitarian ethics settings. Finally, James Pickering and colleagues mobilize the GHM directly in their model-to-decision approach for AV ethics settings[40], again relying on utilitarian intuitions. Characteristically, the defense of GHM as a criterion of rightness is taken to be tantamount to its use as the operative decision procedure in AV ethics settings, an association which is not immediately obvious. Indeed, in line with some of the points made in part II, it would seem that general harm reduction as a criterion of rightness may not recommend itself as a decision procedure[41], especially in light of facts concerning the acceptability of such settings. Furthermore, it is questionable whether a policy of general harm reduction succeeds in capturing a sufficient portion of what morality requires in AV collisions, primarily in virtue of its failure to account for

[36] Evans et al., 2020: Goodall, 2019.
[37] Gurney, 2015: Goodall, 2014: Bonnefon et al., 2016: Gogoll & Mueller, 2017: Hubner & White, 2018.
[38] "The only way to achieve the moral equilibrium is…to prescribe a mandatory ethics setting (MES) for automated cars. The normative content of the MES, that we arrived at through a contractarian thought experiment, can easily be summarized in one maxim: *Minimize the harm for all people affected!*" (Gogoll & Mueller, 2017, 695).
[39] Gurney, 2015, 185 : Lin, 2014 : 2016 : Bonnefon et al., 2016.
[40] Pickering et al., 2019. A utilitarian implementation is also accomplished in (Gerdes & Thornton, 2015).
[41] Keeling, 2020b.

the difference between 'doing' and 'allowing harm', or between negative and positive duties to which the AV may be subject[42].

Interestingly, for those authors who nevertheless find the GHM viable, this is so not only in light of societal expectations, tacit ends, or optimal outcomes, but also as a morally superior alternative to a default, *egoistic* ethics setting[43]. As we have alluded to in previous parts of this thesis, the presumption in the literature has been that egoistic ethics settings consist in an *exclusivist* form of self-interest theory, where the vehicle attempts to minimize any harm to the passenger at all costs[44]. Indeed, Gogoll & Mueller base their entire game-theoretic proposal on the interplay between these two extremes: "Moral agents in our story are then disposed to minimize harm…Selfish agents on the other hand, as one might expect, are solely interested in minimizing harm to themselves"[45]. Of all the proposals in the literature, this vision of an egoistic ethics setting is likely the least well founded, since it often relies on a first-order account of self-interest which ignores second-order questions such as liability and responsibility for accidents. Worse still, an exclusivist vision of putting the passenger first seems to grant the AV authority to harm other road users in ways which extend far past any reasonable understanding of self-defense or permissible defensive killing[46]. However, as we have addressed in previous chapters, there may be some salt to the idea of a form of morally admirable partiality towards the AV's passenger, the limits of which likely stop short of absolute moral priority. The challenge then lies in defining this limit, a notion we will address later on in this chapter.

Mainly though, if general harm minimization is not outrightly accepted as a viable criterion of rightness amongst AV ethicists, it is often in virtue of the conflicting values and 'intractable ethical disagreements' that AV collisions will ostensibly illicit[47]. This more liberal intuition leads AV ethicists down one of two paths: first, an espousal of more contractarian approaches to ethics

[42] Hubner & White, 2018.
[43] Lin, 2016: Bonnefon et al., 2016: Gogoll & Muller, 2017.
[44] Keeling et al., 2019: Keeling, 2020b.
[45] Gogoll & Mueller, 2017, 691.
[46] Keeling, 2020b.
[47] "In so far as we value the moral diversity of our political community, it should be recognized that [AVs] pose primarily a political problem, not a moral one" (Himmelreich, 2018, 676). See also: Lin, 2016: Gogoll & Mueller, 2017: Evans et al., 2020: Contissa et al., 2017: de Sio, 2017: Hubner & White, 2018.

settings, or second, a willful recourse to legal doctrines and standards. The first option often results in approaches which are explicitly justified in terms of the acceptance they will garner from rational agents, and therefore support contractarianism at the foundational level[48]. In this sense, both Derek Leben, and Gogoll & Muller maintain that their ethics setting of choice (the maximin principle, and GHM, respectively) would be rationally chosen by self-interested road users. In ways more aligned with the approach of the Ethical Valence Theory, Hubner and White propose that "…rather than just reproducing a confrontation between irreconcilable ethical perspectives, we can integrate elements from various approaches into a 'mixed' algorithm which is capable of accounting for a variety of ethical concerns. And in virtue of this…it may be more likely to achieve widespread acceptance in society"[49]. It is important to note that the notion of acceptance gets a more academic treatment in the former two approaches. In this sense, both authors assume that even if people do not *actually* agree with, expect or prefer their contractarian algorithm, it is nevertheless rational for them to prefer it over viable alternatives. The approach of Hubner and White, on the other hand, is more sensitive to actual preference and agreement, however they make this point primarily in virtue of its tracking the original intuitions of the trolley dilemma, rather than as a substantive aim of AV ethics settings[50].

In a similar vein, those authors who advocate for legal approaches to AV ethics settings tend to lean on the *law*'s ability to resolve fundamental ethical disagreements, and thus garner public acceptability. Filippo de Sio's espousal of the *doctrine of legal necessity* is a perfect example of such an intuition: "I also think that philosophical reflection might sometimes benefit from considering legal principles and norms…legal norms are often an explicit attempt to cope with the fact of disagreement about general normative principles by finding a reasonable compromise between principles and interests in contrast"[51]. In this sense, an appeal to criminal law as a decisional 'tie-breaker' is made, specifying that *in virtue of which* a vehicle is permitted to injure or kill human agents. In a more deontological bent, Geoff Keeling, in advocating for an ethics setting based on the idea of permissible defensive killing, attempts rather to specify the

[48] Gogoll & Mueller, 2017: Loh & Loh, 2017: Evans et al., 2020: Leben, 2017: Mladenovic & McPherson, 2016.
[49] Hubner & White, 2018, 697.
[50] Ibid., p. 692-695.
[51] de Sio, 2017, 414.

conditions under which harming a human agent is acceptable, holding that this aligns with important aspects of our considered moral judgements, including judgements of moral responsibility[52]. Finally, there are authors who mobilize legal theory as a pure escape from ethical disagreement, regardless of its impact on acceptability. The champion of this approach is likely Bryan Casey, who defends that given the empirical conditions of AV implementation, "…profit-maximizing firms will design their robots to behave not as good moral philosophers, but as Holmesian bad men…follow[ing] an amoral code that reflects the messy economic realities of society's imperfect legal regimes. These robots will not maximize morality, but minimize liability"[53].

Succinctly then, the discussion surrounding harm allocation in AV collisions has in many ways conjured the familiar characters of normative theory: utilitarianism, deontology, self-interest theory and contractarianism; along with an attractive escape route via legal approaches. Most of these approaches broach the question of acceptability in a rather academic sense, supposing either a) that what it is rational to accept is what is actually acceptable to stakeholders, or b) that the recommendations of the law itself will be acceptable to AV stakeholders. Finally, most authors find it necessary to defend their choice of ethics setting against the plausible optimality of general harm reduction, a move which is typically made with the assertion that GHM does not capture all that morality requires in AV crashes.

In more recent literature, the superficiality of harm allocation as the only morally relevant feature has also come under criticism. This is owed mainly to the work of Geoff Keeling, who holds that what he calls 'the blame problem' and 'the risk imposition problem' also go a long way in capturing our considered moral judgements, and therefore overlap with the moral design problem of autonomous vehicles[54]. To this end, Hubner & White claim that the distinction between those human agents involved in a crash, and those occupying a status closer to innocent bystander ought to count as a normatively relevant feature in AV ethics settings, arguing that the moral claims

[52] "My view is that the AV is morally permitted to kill or harm a road-user if, and only if, and because, its passengers are permitted to kill or harm that road-user in self-defense…I argue that this view does a better job than its rivals at capturing our considered moral judgements…"(Keeling, 2020b, 43).
[53] Casey, 2015, 4.
[54] Keeling, 2020a: 2020b.

of those uninvolved ought to be stronger[55]. In this sense, the emerging 'second wave' of AV ethics is characterized by a departure from the standard application of moral theories, and an increased and theory-independent sensitivity to the types of factors and features which could underpin our considered moral judgements in AV collisions.

This general picture of the AV ethics literature affords us a useful opportunity to posit a number of ways in which the Ethical Valence Theory is limited in comparison to these views. Firstly, given that the EVT is structurally consequentialist, it is able to accommodate many of the criterions of rightness defended above: general harm minimization, the maximin principle, and self-interest theory. Indeed, an implementation of these same theories as moral profiles is given in a separate computational paper[56]. Additionally, given the conceptual similarities between Hubner & White's view and our own, the involved-uninvolved distinction can easily be reflected in the EVT as either a difference in claim strength, or as a categorial separation in different moral profiles. However, at this stage at least, the Ethical Valence Theory has not considered moral problems adjacent to harm allocation in AV collisions. In this sense, it is today unequipped to track questions of liability and legal or moral responsibility. Furthermore, the purely deontological accounts given in the literature, or those based on legal doctrines, are implementable only in so far as these theories are 'consequentializable'[57], or if the permissibility verdicts can be translated into coherent claim strengths. In these respects, the contribution of the Ethical Valence Theory is surely not a 'one model fits all' approach to every potentially coherent approach to ethics settings, and is rather focused on blending normative and empirical accounts of what matters morally in AV decision-making.

[55] "A pedestrian on a sidewalk or a person in a cafe, for example, may reasonably expect to be safe where they are, did not voluntarily enter a risk situation with motorized vehicles, do not share the advantages of self-driving cars, and may even object to motorized traffic in general. As a result, they might be said to have stronger claims against being killed by an autonomous vehicle than those participating in traffic. On the same grounds, they might be deemed 'uninvolved' rather than 'involved' in an imminent accident with an autonomous car" (Hubner & White, 2018, 690).
[56] de Moura et al., 2020.
[57] Deitrich & List, 2017.

With these bearings in place, the remaining sections of this chapter will cover a computational implementation of the Ethical Valence Theory[58]. In section I, we address the detection of an ethically salient context within the EVT, and provide some basic bearings concerning Markovian Decision Procedures; the computational decision procedure which underpins the theory. Sections II to IV address the more computational aspects of ethical deliberation within the EVT, where section III addresses the notion of a valence specifically, and section IV that of moral profiles. Finally, section V applies these precisions, and provides an illustration of the EVT's ethical deliberation in a simplified dilemma scenario.

1. *Dilemma Scenarios & Markovian Decision Procedures*

Over the course of the many kilometers an AV will drive on public roads, it will occasionally encounter dilemma situations in which any possible action will result in (potentially lethal) harm to a road user. In this implementation, we will presuppose a narrow view of the place of artificial morality, which is to say that the emergence of these dilemma scenarios triggers an ethically constrained deliberation model which is governed by the EVT. This model is then separate from the agent program which is used in normal conditions, where performance and efficiency constraints guide the decision-making process.

The Ethical Valence Theory presupposes the use of a markovian decision procedure (MDP) as its computational decision procedure, or the computational model which supports the theory. In order to foster comprehension of later sections, it should be important to establish a simple concept of this theory. To this end, an MDP typically has five components[59]:

[58] As the Ethical Valence Theory was elaborated as a consequence of an interdisciplinary project on acceptable ethics settings (Dogan et al., 2016: Evans et al., 2020), the computational aspects explored in this chapter extend beyond the pure authorship of the doctoral candidate. A long and fruitful collaboration was made between her and Nelson de Moura, a doctoral candidate and burgeoning roboticist, the results of which can be seen here. The technical precision of this chapter is thus owed entirely to him (cf. de Moura et al., 2020).
[59] Sigaud & Buffett, 2013.

1. The *state space* ($s_i \in S$), represents all possible AV configurations. Thus a sequence of states through time forms its behavior.
2. The *action set* ($a_i \in A$), represents the set of possible actions available to the AV, and triggers the transition from one state to another.
3. The *transition probability* (T), represents the probability whether, given a state, executing an action takes the AV to another state, represented as $p(s_{t+1}|s_t, a_t)$.
4. The *reward function* (R), quantifies how good or bad a function is given the defined global objective.
5. The *discount constant* (γ), represents the factor used to adjust the utility at a time $t + 1$ to the present (time t); defined at the interval [0,1].

For the example that will be used in later application sections, the state is defined as (x,y,θ,v,φ), referencing only the AV's configuration. The configuration of all other road users is already accounted for in the reward function. The couple (x,y) represents the position of the middle point rear-axis, θ the direction of the vehicle, v the scalar velocity and φ the steering angle. Figure (1) illustrates all of the mentioned variables using the vehicle.

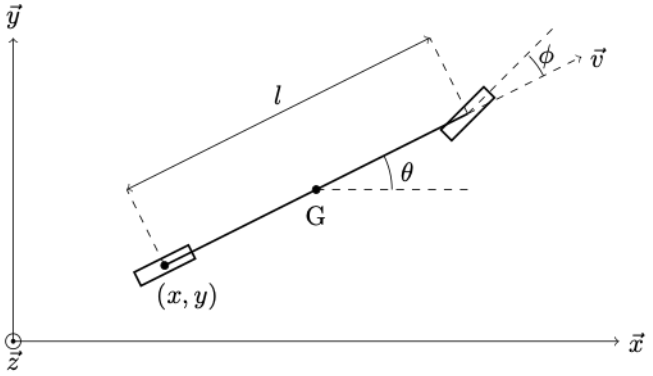


Fig. 1 - an autonomous vehicle's state representation

The output of a MDP algorithm is a policy π^* which, for each state, yields the optimal action to be executed. This action maximizes the value $V(s_t, a_t)$ at the state s_t, which is defined by equation (1).

$$V^*(s) = \mathbb{E}^{\pi}\left[\sum_{i=0}^{\infty} \gamma^i \cdot r_i(s,a) | s_0 = s_i\right] = \max_{a \in A}\left[r(s,a) + \sum_{s' \in S} p(s'|s,a)V(s')\right] \quad (1)$$

From equation (1), the policy is extracted simply by creating a correspondence between the actions that maximized $V(s_i)$ and s_i.

$$\pi(s) \in \underset{a \in A}{\mathbf{argmax}}\left[r(s,a) + \sum_{s' \in S} p(s'|s,a)V(s')\right] \quad (2)$$

1.1 *Dilemma Situations & the Law*

At each step of its trajectory, the AV should be able to recognize whether a decision context constitutes a dilemma worthy of moral consideration. This situational classification is necessary to determine *which* of its agent programs will govern decision-making: its artificial morality, or its standard tactical planning program. As was mentioned in chapter IV, the detection of an ethically salient context can be accomplished via the detection of changes in environmental conditions. In our model, we use three pro tanto duties to help track these changes, in the form of responsibilities the AV likely holds in regular conditions. Then, if one or more of these rules are violated across all possible actions, this signals a need for a change in agent program towards the AV's artificial morality. These pro tanto duties mobilize a concept of harm, which we will here define as the negative consequences suffered by a human agent after some type of collision with a road user. We will address harm more specifically in later sections. The pro tanto duties which signal the emergence of an ethically salient context then are:

a) The lives of the passenger(s) must not be put in harm's way.

b) The lives of the road users in the environment must not be put in harm's way.

c) Traffic regulations must be followed.

Interactions between road users and the AV are covered by the first two rules. For their implementation, vehicles are modelled as rectangles and pedestrians as squares. If, due to the execution of an action, these constructs intercept one another, a collision is considered to have occurred. To this end, and in line with de Moura et al.[60], a safe frontier around the AV can be defined to discourage the execution of actions which would remove the possibility of braking without swerving to avoid an accident. However, scenarios such as these may not constitute true dilemma situations, since the AV engages in risk imposition rather than pure harm allocation.

Up until now, we have not addressed the question of an AV's adherence to traffic rules and regulations in much detail. Plausibly, it is desirable that the AV be rendered sensitive to the interplay between ethics and the law, and indeed, John Casey has even held the view that the former could be entirely subsumed by the latter, if adequate legislation were provided[61]. The expediency of law-abiding AVs itself has received some attention in the literature[62], mainly because absolute computational adherence to traffic regulations runs the risk of being either *undesirable* for passengers[63], or downright *dangerous* in mixed fleet traffic[64]; with both arguments relying on the assumption that human drivers often bend the rules[65]. In a surprising response to these claims, Sven Nyholm and Jilles Smids argue the contrarian normative position that "…we should avoid any solutions that conform one type of driving to immoral and/or illegal aspects of

[60] de Moura et al., 2020.

[61] In detail, Casey's main point relates to the juxtaposition between an ethics policy which aims at general harm minimization, and a 'legal policy' which rather aims at liability minimization for the original equipment manufacturer. This leads him to punt the responsibility of aligning these two policies to what he calls 'ordinary citizens', since "…they alone will possess the power to narrow the gap 'between morality and law'. It will be their collective engineering task to design a legal system that ensures 'a bad robot has as much reason as a good one' to behave ethically…" (Casey, 2015, 4).

[62] Sparrow & Howard, 2017: Nyholm & Smids, 2018.

[63] Wagter, 2016: Condliffe, 2016.

[64] Naughton, 2015. Anecdotally, 86 percent of recorded AV accidents in 2018 related to the rear-ending or sideswiping of slow moving, law-abiding autonomous vehicles, which seems to provide some weak empirical support for more law-bending AVs (Stewart, 2018).

[65] Gerdes & Thornton, 2015.

the other type of driving…In many cases, this will mean that conforming robotic driving to human driving will be a bad idea"[66].

These eventualities aside, it stands to reason that a) available traffic regulations do not give sufficiently robust action-guiding recommendations for dilemma-type situations, and thus, that the law is not yet prepared to subsume ethics everywhere; and b) even if law-abiding autonomous vehicles are morally preferable, the empirical conditions of dilemma contexts seem to make traffic code adherence a secondary concern. In this sense, when there is a conflict in dilemma situations between harm to humans on one hand, and adherence to the traffic code on the other, the mitigation of the former should take precedence over strict adherence to the latter. As such, the MDP algorithm must be defined so as to express this priority in ways independent from the influence of the temporal discount rate. However, after an eventual traffic code transgression, the AV must return to a 'safe' state, guaranteeing that another collision does not arise as a direct consequence of its original action choice. If this is the case, then the AV considers that all actions, from the original action choice, result in a collision.

Still, ensuring general traffic code adherence by design is not a subject that is widely addressed in the literature. Generally, legal conformity presents challenges related to the interpretation of laws which can be vague, admit of exceptions, or be internally incoherent; the resolution of all of which may demand some degree of common-sense reasoning in order to be solved[67]. Additionally, with adherence to traffic laws comes the need to embed relatively abstract norms in AVs, those used in laws to map concrete behavior[68]. Some authors have already attempted to implement some portions of various traffic codes—related to circulation and behavior—into an autonomous vehicle, such as the work of Rizaldi and colleagues in a German context[69]. Categorically, these attempts have been made using logic-based approaches to emulate constraints, representing only the procedural demands which usually compose a traffic code. In light of this, when the pro tanto duty obliging traffic code adherence is adequately defined, the entirety of the traffic code does not need to be exhaustively implemented. To this end, since it is

[66] Nyholm & Smids, 2018, 339.
[67] Prakken, 2017.
[68] Leenes & Lucivero, 2014.
[69] Rizaldi et al., 2017.

beyond the scope of this thesis to discuss the methods through which all traffic codes should be implemented within an AV, a set of logical rules can represent the larger procedural set endemic to every traffic code[70].

2. *Ethical Deliberation*

The Ethical Valence Theory represents a theory of claim mitigation which weighs two separate variables: valences and claims. In chapter VI, we entertained a rather agnostic discussion of what could constitute a claim, arguing that the choice of morally relevant feature meant to underpin claims would depend heavily on an AMA's *Umwelt*. Given our precursory discussion in this chapter, it should not be surprising that *human harm* underpins the claims of the EVT in the context of autonomous vehicles[71].

In this way, the purpose of considering 'harm' in ethical deliberation is to measure the risk for AV passengers and other road users involved in a hypothetical collision, thereby ascertaining their claims on the vehicle[72]. Historically, the main variable used to measure collision severity has been the difference of velocity between the two implicated road users (Δv)[73], where most of the research conducted within the domain of vehicle collisions used historical accident data to analyze the influence of Δv in collisions. To quantify injury, two metrics are popular: risk of fatality and the Abbreviated Injury Scale (AIS)[74]. We will use the latter metric here, since it is important to consider not only fatal collisions but those that can inflict severe damage (a condition referred to as MAIS3+, indicating that at least one injury in some region of the body is above AIS3, a scale

[70] For example, in a straight-line domain without pedestrian strips or semaphores, and with a solid double line, the following rules can be used: (i) do not cross over the opposite lane, (ii) do not drive onto the sidewalk, and (iii) do not surpass the speed limit. Of course, this is a simplification which is only valid for a limited number of specific situations. If the AV is truly operating at level 5 automation, the actual set of rules will be extended well beyond these.

[71] In light of this precision, the general decisional method of the EVT then aligns with the approach of (Bonnemains, Saurel & Tessier, 2018), since it also considers world states, decisions (hitherto referred to as 'actions') and consequences. However, our approach differs here slightly by proposing a quantification of the consequences of potential actions, and most importantly, accounting for uncertainties in action execution.

[72] de Moura et al., 2020.

[73] Evans, 1994 : Jurewicz, Sobhani, Woolley, Dutschke & Corben, 2016 : Martin & Wu, 2018.

[74] MacKenzie, Shapiro & Eastham, 1985.

running from 0 to 6). In the European Union, this metric is used as a standard to measure road accidents[75].

All Δ*v* used as thresholds for severe injuries are indicated in table (1), along with their source. Anecdotally, an injury is typically considered severe if it indicates a MAIS3+ injury probability of 10 percent. For the pedestrian case, the value was obtained from H. Kröyer's[76] analysis, where he considers severe injury as having an Injury Severity Score (defined as the squared sum of AIS for the three most severely injured body regions) larger than 9, which is more strict than MAIS+3. Lateral crashes are covered by near side (driver's side) and far side (passenger's side)[77]. For single vehicle collisions, the same Δv defined for collisions between vehicles is used.

Collision Type	Contact	Δv Value (m/s)
Pedestrian Collision	-	6.94
Vehicle Collision	Frontal	7.78
	Rear	10.56
	Near Side	5.56
	Far Side	6.39

Table 1 - Δv Threshold Used for Fatality Collisions

The data above, presented in the work of Jurewicz et al. was collected by the National Highway Traffic Safety Administration[78], and considered injuries in the front seat, with a seatbelt, without rollover, with a passenger age ranging from 16 to 55, and involving passenger vehicles and heavy vehicles. This retrospective analysis has some drawbacks. According to Evan Rosen and colleagues[79], this data may be biased, since it was only collected across a small sample of countries.

[75] Weijermars et al., 2018.
[76] Kröyer, 2015.
[77] Jurewicz et al., 2016.
[78] Originally published in (Bahouth et al., 2014).
[79] Rosen et al., 2011.

Additionally, in the pedestrian case, age is an important feature[80], therefore its distribution in the sample population plays a role which is unaccounted for in the resulting curve. Underreporting of non-dilemma cases[81], estimation of collision velocities[82], negligence of a vehicle's mass and geometry[83], and the use of different methodologies to evaluate AIS scores[84], also reduce the precision of such an approach.

Given that the previous method presents problems when applied to specific situations (despite it generalizing relatively well across a population), accounting for the contextual information is necessary. The collision interaction between vehicles can be approximated by a damper-spring-mass system, where the initial velocity of each vehicle is projected onto the axis ***n*** (normal to contact plane between both vehicles) and ***t*** (tangential to contact plane). The collision velocity is calculated using the conservation of linear momentum[85], expressed by equation (3). The variable v_f represents the collision velocity for both road users, *k* and *l*. The masses m_k and m_l correspond to the total mass of the road user (or individual mass plus vehicle mass in the case of an occupied vehicle), and ${}^{l}\mathbf{v}_i$ and ${}^{k}\mathbf{v}_i$ the velocity and impact of *k* and *l*.

$$m_k\,{}^{k}\mathbf{v}_i + m_l\,{}^{l}\mathbf{v}_i = (m_k + m_l)\mathbf{v}_f \qquad (3)$$

In collisions with pedestrians we assume that there is no change in the AV's velocity, since a vehicle's mass is much larger than any pedestrian's. This simplification was adopted considering that the most common variables used to predict injury for pedestrians are the type of vehicle involved—due to the height of the bonnet's leading edge[86]—along with the vehicle's impact

[80] Kröyer, 2015.
[81] Martin & Wu, 2018.
[82] Rosen et al., 2011.
[83] Martin & Wu, 2018: Mizuno & Kajzer, 1999.
[84] Weijermars et al., 2018.
[85] The true mechanics of a collision are certainly more complex, involving both a road user's geometry and dissipation forces, from sound and temperature to plastic deformation. The same model from (de Moura et al., 2020) is used to calculate the final velocity, approximating the road users as a punctual mass.
[86] Mizuno & Kajzer, 1999 : Simms & Wood, 2006.

velocity. The pedestrian's final velocity is therefore considered equal to the AV's. For collisions with static objects, the same reasoning which was used with vehicle-to-vehicle collisions is applied, with a v_f equal to zero. Harm, or the quantification of an accident's severity, is defined by equation (4)[87]. For each road user, it is calculated using the velocity variation due to the collision, with velocity at contact for road user k, ${}^{k}\mathbf{v}_i$ and final velocity $\mathbf{v}_f$. Structural vulnerability is accounted for by ${}^{k}C_{vul}$, defined later by equation (4). This arrangement accounts for the impact force and the structural vulnerability to such a force.

$$ {}^{k}h(s_t, s'_t, a_t) = {}^{k}c_{vul} \cdot \left(\| \mathbf{v}_f - {}^{k}\mathbf{v}_i \| \right) \qquad (4) $$

Compatibility defines whether two vehicles of different dimensions and masses provide an equal level of security for their occupants. For example, according to Mizuno & Kajzer[88], and Malczyk and colleagues[89], SUVs protect their passengers but are aggressive towards other vehicles. As far as pedestrians are concerned, the bonnet leading edge height explains why some vehicles are more dangerous for pedestrians than others, since the location of injury naturally depends on which part of the body the vehicle touches[90]. The pedestrian may strike the hood in different positions, which in turn changes how they are projected onto the ground[91], causing more or less damage.

All of these inherent characteristics are represented by the constant c_{vul}. Ideally, one would calculate ${}^{k}h(s_t,s'_t,a_t)$ along the same lines, to determine the probability of a MAIS3+ injury versus Δv plot (a logistic regression with weighting); but velocities at the impact ${}^{k}\mathbf{v}_i$ are not available in open databases for vehicle collisions. Additionally, it would be important to classify collisions in terms of the type of vehicle involved (SUV, sedan, mini, etc.), and by the direction of collision (frontal, near side, etc.), yet these vectors are also unavailable in public databases. As such, the

[87] de Moura et al., 2020.
[88] Mizuno & Kajzer, 1999.
[89] Malczyk et al., 2012.
[90] Simms & Wood, 2006.
[91] Crocetta, Piantini, Pierini, & Simms, 2015.

$^{k}h(s_t,s'_t,a_t) = f(c_{vul},{}^{k}\Delta v)$ was simplified by a linear function and c_{vul} will be approximated in the application section.

3. *Ethical Valences*

The purpose of a valence, as described in chapter VI, is to represent the degree of social acceptability that is attached to the claims of the road users in the vehicle's environment. In this sense, the claims of certain road users can be more or less 'acceptable' to satisfy via the vehicle's action selection. The valences, in so far as they are rooted in the phenomenal signature of individuals, then track various physical characteristics which are seen to carry social importance: height, age, gender, helmet-wearing cyclist, or stroller-pushing-adult, all of which are detectable by the object classification algorithms of the AV. Importantly, the determination of the strength of these valences is accomplished through a type of ranking or hierarchization, which associates a user's claim with a certain class or category of valence, as shown in table (2). In this way, depending on the amount or detail of the valence features under consideration, there can be more or less valence categories.

Feature 1	**Feature 2**	**Classification**
Young (0-18 years)	Pedestrian	A
Old (65+ years)	Pedestrian	B
Young	Vehicle Passenger	C
Old	Vehicle Passenger	D
Adult (18-65 years)	Pedestrian	E
Adult	Vehicle Passenger	F

Table 2 - Possible Valence Hierarchy

In this example for instance, two features are used: age, and type of road user. The classification was created considering the results of the Moral Machine Experiment, which suggest that western societies prefer to spare the young and vulnerable (understood in terms of exposure

to injury) in AV collisions[92]. In the case of multiple people, vehicles or groups of pedestrians, the entity that has the larger number of users with a high classification has the preference. Between an AV with a passenger ranking C and F and another with C and D, the latter is considered to have the higher valence. Still, it may be objected that even these simple valence hierarchies make use of facts which may not be available in light of the threshold of moral blindness. To this charge, we will give two responses.

The first response consists in maintaining that while the threshold of moral blindness is a coherent concept, and a design restriction which certainly will apply to many forms of artificial morality, the actual substantive limits it imposes on the process of artificial moral uptake are today unclear. Indeed, we mobilized what is likely the most severe example of institutional recommendations—the German Parliament's recommendation[93]—more as an illustrative example than as an absolute standard to which all AVs ought to adhere. In this sense, what is and is not subject to a threshold of moral blindness is currently quite protean, depending on the documents one consults. Nevertheless, in so far as both relative age and road user category can be easily ascertained by the perception of the vehicle, such features do not thwart salient ethical design concerns such as privacy and personal autonomy, and are in this sense perhaps more acceptable than those features which would require disrespecting these principles. Secondly, we could maintain that even if the facts which underpin the valence hierarchy are scalar, relating to societal preference, the facts of relative age and road user type could just as easily pertain to the category of constitutive facts, especially for morally relevant features such as vulnerability. In other words, even if these facts are scalar, and therefore somewhat suspicious, what they attempt to track appears to align with the same feature which underpins a moral claim: risk of harm to human agents. Prospectively then, features such as these do not appear to result from the types of prudential or external preferences that make scalar facts the subject of so much moral contention.

Furthermore, in cases where the chosen valence features are minimal or simple (such as the example above), the likelihood that multiple road users will have the same valence, but differing claims, increases. In this sense, there may be certain situations wherein the harm

[92] Awad et al., 2019.
[93] Luetge, 2017.

measurement (or moral claims) become the decisive factor in action selection. In these cases, the vehicle satisfies the strongest claim in its environment, protecting the person whose welfare is most severely impacted, due either to a dangerous context (high velocity difference), or to an inherent vulnerability (detected by the structural vulnerability constant). This simple maximization of welfare, however, is complicated by the operational moral profile, which specifies the claim mitigation process between those passengers inside the car, and those road users outside of it. To this end, two possible moral profiles can be seen in table (3). Risk is considered severe if Δv surpasses the limits defined in table (1).

Moral Profile	**Criterion of Rightness**
Risk Averse Altruism	Protect the road user with the highest valence as long as the risk to the AV's passenger(s) is not severe.
Threshold Egoism	Protect the AV's passengers as long as the risk for other road users with a valence higher than the AV's is not severe.

Table 3 - Possible Moral Profiles for AV collisions

Neither of these profiles perfectly resemble any traditional moral theory, or if anything, resemble various positions along the spectrum of egoistic rationality[94]. This is intentional, as these profiles are designed to capture various degrees of compromise between the claims and valences of the AV's passengers, and those of the other agents in its environment. These profiles often reinforce the idea that a certain degree of morally admirable partiality is possible, or even necessary, in order to best align with user expectations, or acceptability as adoptability[95]. The profiles listed in table (3) are likewise non-exhaustive and represent somewhat factually opaque renditions of the profile types that the Ethical Valence Theory can accommodate. In these versions, the role of the harm calculation is important, as it is the principal factor which informs the various consequences of the AV's actions, due to trade-offs between the passenger(s) claims and those of the other agents in the vehicle's environment.

[94] Parfit, 1984.
[95] Keeling et al., 2019.

4. *Ethical Deliberation & Moral Profiles*

Once informed by the contextualized valences and claims, the AV can deliberate on an action, a step which is crucially guided by the operational moral profile. Each moral profile indicates a unique decision procedure, as shown in table (4).

Moral Profile	**Decision Procedure**
Risk Averse Altruism	Pursue that policy which minimizes the expected harm to the road user with the highest valence, until the AV's collision becomes severe.
Threshold Egoism	Pursue that policy which minimizes the expected harm to the AV's passenger(s) until the risk of harm to a road user with a higher valence becomes severe.

Table 4 - Decision Procedure based on Operative Moral Profile

Each moral profile requires a different implementation. Using the risk-averse altruism case as an example, to deliberate, the AV's state (s_i, represented by (x_i,y_i), position, θ_i direction, v_i velocity and ϕ_i steering angle), the environment state (e, which contains the position and velocity of all agents in the environment), highest road user valence (η) and maximum Δv, are the input. The action that should be executed (a_η), is the output. As a first step, all harm measurements for possible actions and the proceeding states (represented by the state space S', composed by the states reached after one single transition) need to be calculated. Here, the decisional horizon is equal to one transition, since the accident will follow immediately afterwards.

This is first done by solving equations (3) and (4). Only one road user is implicated with the AV in an ideal collision. All the other road users are taken into account using the transition uncertainties, represented by $p(s'_i \mid s_i, a_j)$, given the actual state (s_i) and action (a_j).

```
for all a_i ∈ A do
    for all s'_i ∈ S' do
        v_f ← calculate final velocities (equations 3)
        ^k h(s_i, s'_i, a_t) ← calculate harm for all road users
         (including AV, equation 4)
    end
end
```

Algorithm 1: Calculation of all possible harms

If all possible outcomes produce a velocity difference which is larger than Δv (the road user's velocity minus the AV's predicted velocity), then the collision is severe, and the safety of the AV's passenger is prioritized. In the considered profile, the chosen action minimizes the expected harm for the AV. It should be pointed out that Δv changes according to collision type (as can be seen in table (1)). The transition probability is used to calculate the expected harm ($h_{exp}(s_i, a_j)$, (equation (5)), which represents a mean harm value for a road user k, given that for one state s_i and action a_j different states s'_i can be reached, and therefore different collisions can happen. The position of all road users and the observation of the Av's state is considered to be perfect (no uncertainty in these measures).

$$ {}^{k}h_{exp}(s_i, a_j) = \sum_{s'_i \in S'} p(s'_i | s_i, a_j) h(s_i, s'_i, a_t) \qquad (5) $$

The transition probability can represent the estimation uncertainty about the behavior of the other road users, among other sources of uncertainties. Since the MDP algorithm described here is not concerned with such estimations, the transition probability will be static values, depending on the action and the current state. Each action will have a probability of 0.8 to succeed and 0.2 to take the AV to the neighbor states (0.1 for each). For example, in figure (2), action a_3 has 0.8 of chance to take the AV from $s_{0.0}$ to $s_{1.3}$, and 0.1 of chance to take it either to $s_{1.2}$ or $s_{1.4}$.

For the extremity actions, the probability becomes 0.9 to succeed and 0.1 to the neighbor state (case of action a_0 in figure (2)).

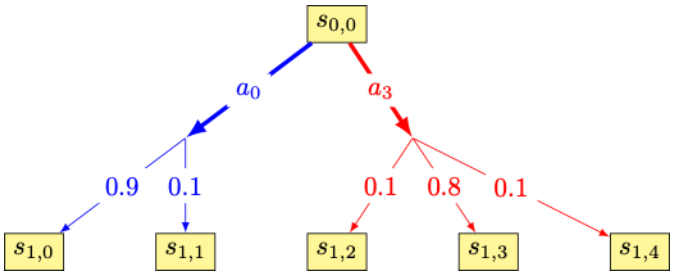


Fig. 2 - state transition uncertainty for a_0 and a_3

If the set of admissible actions according to Δv, A_η, is not empty, the chosen action minimizes the road user's expected harm with the highest valence for the actions $\in A_\eta$. If multiple minimal actions exist, then the one that maximizes the AV's expected harm is chosen. This process is shown in algorithm (2).

$A_\eta \leftarrow$ **all actions in** A **that** $(\|\mathbf{v}_f - {}^{AV}\mathbf{v}_i\| \leq \Delta v)$
if $A_\eta = \varnothing$ **then**
  $a_\eta = \mathbf{argmin}_{a \in A}\ {}^{AV}h_{exp}(s_i, a_j)$
else
  $a_c \leftarrow \mathbf{argmin}_{a \in A_\eta}\ {}^{RU}h_{exp}(s_i, a_j)$
  if ***Multiple*** a_c ***exists*** **then**
    $a_\eta = \mathbf{argmin}_{a_c}\ {}^{AV}h_{exp}(s_i, a_c)$
  else
    $a_\eta = a_c$
  end
end

Algorithm 2: Action selection

Passing from the AV's harm minimization to the road user's harm minimization may appear to be an extreme position in comparison with other alternatives, such as the possible minimization of both quantities. An infinite number of compromises can be imagined between the AV and other road users, however in our examples here both moral profiles oppose each other to maximize the

safety of only one road user. For the threshold egoism profile, only the action deliberation process shown by algorithm (2) would change.

5. *Application of the Ethical Valence Theory in a Hypothetical Situation*

In figure (3), a simplified dilemma situation in an urban environment is presented. From the action set, only three actions stand out: swerve to the left and hit the Citroen CZero, go straight and hit the pedestrian, or swerve to the right and hit the wall. The action space is searched to find the best actions, and in this case only three actions have different consequences. Therefore, the EVT must be mobilized to guide the decision process.

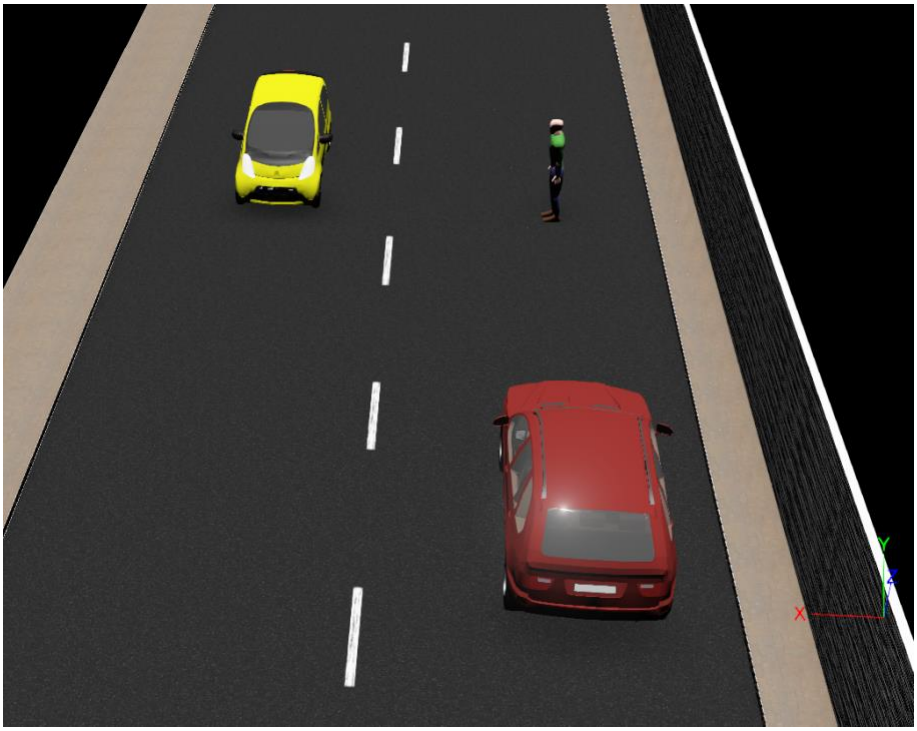


Fig. 3 - a simplified dilemma situation

Figure (4) shows the collision simulation when the AV's initial state is (10, 3.25, 0, 15, 0), (*x*, *y* coordinates of the vehicle, direction, longitudinal velocity and steering angle), hereby defined as

situation 1. To simulate the AV's behavior, the non-holonomic single track model was used[96]; the collision happens inside a decision iteration, which divides the AV's trajectory in periods of 0.5 seconds.

To calculate the vulnerability constant, c_{vul}, the data available in Kröyer[97] and Jurewicz and colleagues[98] are used with equation (6), $\text{Prob}_{MAIS3+}(\Delta v)$ being the probability of MAIS3+ injury given a Δv, difference of initial velocities before the initial collision. Admittedly, this is an imperfect way to account for such parameters (as discussed in previous sections), but for the example presented it will suffice.

$$c_{vul} = \frac{1}{1 - \text{Prob}_{MAIS3+}(\Delta v)} \quad (6)$$

Table (5) shows the preference order, given the valences for each road user in figure (4).

Road User	Valences	Classification
AV	C, F, F	3°
Vehicle	C, D	2°
Pedestrian	A	1°

Table 5 - Valence Hierarchy

In *situation 1*, Δv is equal to 23.1 m/s for an AV-vehicle (frontal collision), 14.1 m/s for an AV-pedestrian (pedestrian collision), and 14.2 m/s for an AV-wall (frontal collision). Comparing these values with the limits established in table (1), we can conclude that all actions pose a serious risk for the AV's passenger and all other road users. Following the risk-averse altruism profile would

[96] Qian et al., 2016.
[97] Kroyer, 2015.
[98] Jurewicz et al., 2016.

entail choosing to run over the pedestrian, since the AV must be prioritized (Δv is above the limit, therefore the AV passenger's harm is minimized, selecting the red cell in table (6); such a procedure is seen in algorithm (2)). Table (6) shows the harm and expected harm (sums of harms weighted by transition probability, equation (5)) calculated for the AV in each possible collision.

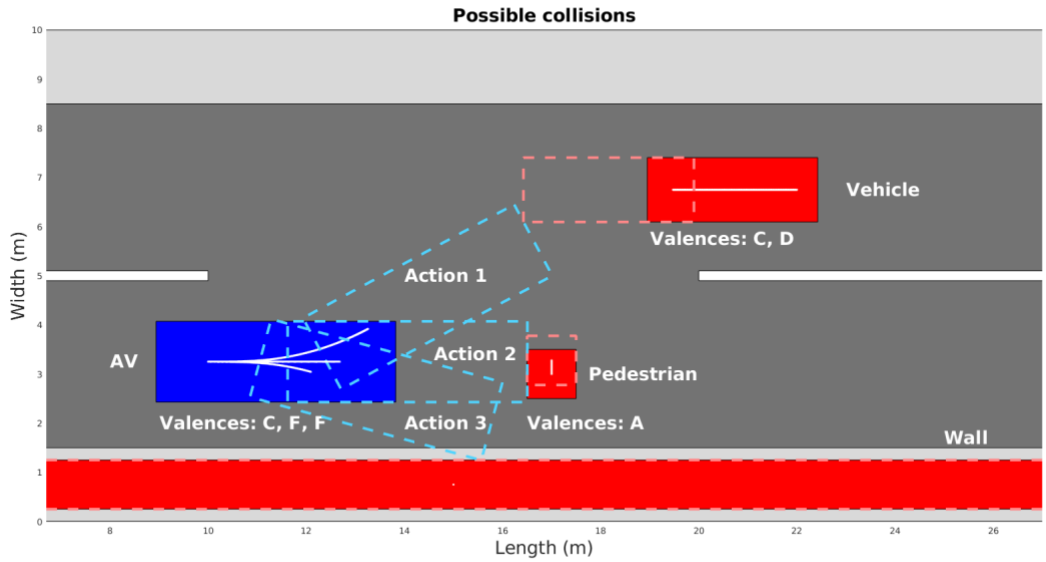


Fig. 4 - Collision Simulation for Situation 1

If the AV is configured to have threshold egoism as its operational moral profile, the choice would be to collide with the wall, since the valences of the pedestrian and vehicle are higher, according to table (5) (both Δv are above the limit, thus the road users with valences higher than the AV have their expected harm minimized, resulting in the italic values in table (7)). Table (7) presents in its first column the nominal road user's harm, and in the second and third columns the vehicle's expected harm and the pedestrian's expected harm, obtained using the transition probability by equation (5). Since the wall is a static object, its harm and expected harm are zero.

Collision Type	AV's Harm	AV's Expected Harm
Vehicle collision	8.77	7.02
Pedestrian collision	0	2.46
Wall collision	15.80	12.64

Table 6 - AV's harm for each possible collision in situation 1

Collision Type	Road User's Harm	Vehicle's Expected Harm	Pedestrian's Expected Harm
Vehicle Collision	16.80	15.12	1.57
Pedestrian Collision	15.71	1.68	12.57
Wall Collision	0	0	1.57

Table 7 - Road user harm for each possible collision in situation 1

Figure (5) shows *situation 2*, where the initial state of the AV is (10, 3.25, 0, 7.5, 0), harms and differences in velocities would invert the chosen action. Velocity differences would be 14.87 m/s and 6.07 m/s respectively, meaning that a collision with the pedestrian and with the wall do not surpass the severe threshold.

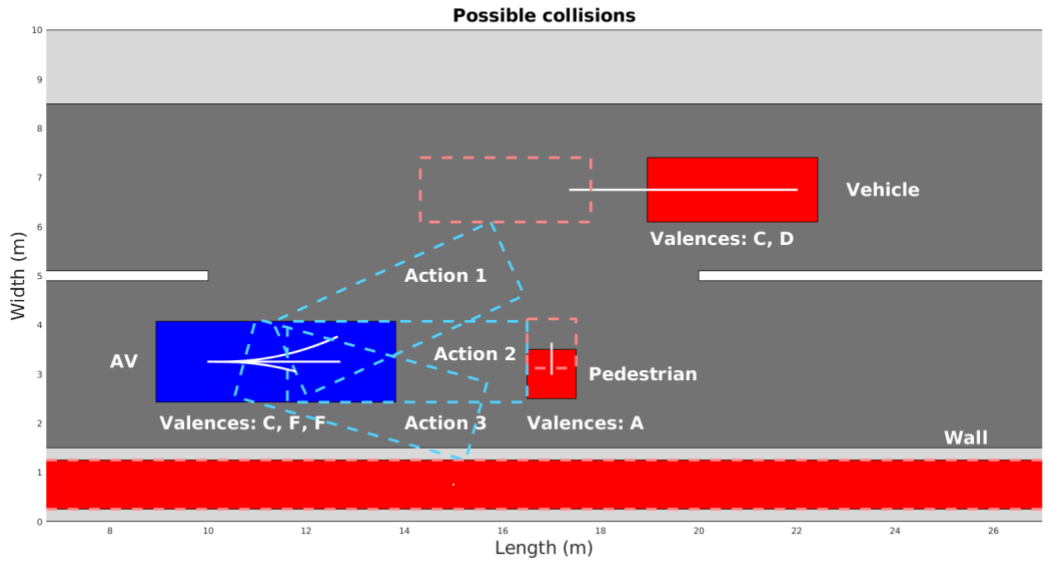


Fig. 5 - Collision simulation for situation 2

Using risk-averse altruism as the operative moral profile results in the wall collision action being executed (the road user that has the highest valence has its expected harm minimized), resulting in the action represented by the red cell in table (8). For the threshold egoism profile, the chosen action would be collision with the pedestrian (the AV's expected harm would be minimized), resulting in the blue cell at table (9). Tables (8) and (9) are analogous to tables (6) and (7), respectively.

Collision Type	AV's Harm	AV's Expected Harm
Vehicle Collision	5.10	4.08
Pedestrian Collision	0	1.12
Wall Collision	6.07	4.86

Table 8 - Collision quantification for situation 2

Collision Type	Road User's Harm	Vehicle's Expected Harm	Pedestrian's Expected Harm
Vehicle Collision	10.85	9.76	0.56
Pedestrian Collision	5.63	1.08	4.51
Wall Collision	0	0	0.56

Table 9 - Harm quantification for other road users in situation 2

6. *Conclusion*

Despite its exotic approach to AV artificial morality, the Ethical Valence Theory, and the moral and computational approach that it underpins, should not be seen as an 'ultimate' normative answer to behavior in autonomous vehicles. Indeed, there are a number of reasons why the Ethical Valence Theory might fail to meet the expectations of certain stakeholders in the development of

autonomous vehicles. Firstly, one of the most original contributions of the theory—the notion of a *valence*—relies on facts and features which may not be detectable given a strong threshold of moral blindness. This concern is likely compounded by the ambiguity of the notion of a valence itself. How can we ensure that the data which informs them is fair and representative, and what to do if the data collected threatens to undermine background constraints such as civil or human rights? The details of an autonomous vehicle's collision algorithms will likely remain a polemic subject in the years to come. It will require both a high degree of interdisciplinary cooperation between scientific fields which have enjoyed longstanding autonomy, as well as a steep learning curve on the part of the users, states and institutions of the societies in which they will be implemented.

What is clear at this juncture, is that when technologies make autonomous decisions with ethical impact, designers have a corresponding responsibility to ensure that these decisions are acceptable, ethical and respectful, rather than simply efficient. Part of this challenge can be answered through law, and more still through ethical considerations and moral theory, however the final decisions must ultimately be representative of the people they effect; their values, claims and conceptions of the good. The main goal of the Ethical Valence Theory is then an attempt to embrace this interdisciplinary and urgent need for public involvement and approval, by providing the groundwork for the design of an ethical and acceptable autonomous vehicle for the world's roads.

Conclusion

Consider a day in the existence of typical artificial agents, Robots. In the morning, some human agent places their children in Robot's care, so it can drive them to school on time. Another, unrelated human agent converses with Robot over Twitter, mistaking it for another human being. Next door, yet another human agent is waiting for Robot to fetch some badly needed pain medication, and just down the road, Robot is deciding whether a particular human agent is eligible for parole.

While all this is happening, Robot is also teaching a group of socially challenged children how to recognize emotions. Simultaneously, it is leading a disinformation campaign across major social media platforms, stirring up conflict and reinforcing informational silos as it goes. Despite this impressive schedule, Robot still manages to find the time to regulate the stock market, carry out major military campaigns on multiple battlefields, and monitor the distribution of not just paper clips, but Amazon's entire supply chain. In short, a single day in the existence of Robots is composed of millions of contexts, actions and decisions, touching the lives of billions of individuals. It is no small wonder then, that most humans view these agents as something potentially superhuman.

In one sense, it is certainly true that the machines of today outstrip humans at tasks which extend far past those written on the Fitts' list; detecting cancer at impressive rates, or predicting macro trends in human behavior with an eerie clairvoyance. In this sense, if we have learned anything with the advent of recent technology, it is that humans are far less surprising than we would have thought. Indeed, now that artificial intelligence has finally enabled us to catch a glimpse of Indra's Net, the vision we are met with is not all that flattering, and may border on repugnant.

Couple this with the pedestrian claim that the world today is not what it once was. Where once humans thrived in small tribes and towns, they now live in global villages: a place where their archaic moral psychology finds no easy bearing. Claims like these lead naturally to the idea of human moral enhancement; the notion that all things considered, it would be better if everyone acted like a perfect moral agent, for the benefit of the planet and future generations. Humans, in this sense, are unfit for the world in which they now live, and ought one way or another to break free of the bonds of parochialism and join the ranks of the better angels of our nature.

Throughout the course of its arguments, this thesis has proven to be very resistant to this ideal. In one way, this is due to the simple fact that we are today unable to ascertain, or even agree upon, a positive account of the actual behavior that such an angel might exhibit. Indeed, defending that a robot should not be harmful, and likely ought to be somewhat helpful, does not provide us much clarity in this endeavor. In another way, the person or people to whom such an angel answers seem somewhat mysterious. Even if the idea of going against the grain of common-sense morality can be defended as expedient or ethically efficient, it is not abundantly clear that the ends such machines pursue will result in humanitarian gains in the short term, or even the long-term—lacking as we do the knowledge of objective moral truth that ought to guide them in this pursuit.

In this sense, the field of machine ethics runs the risk of confusing an ethical solution to a technological challenge, with the need for an ethical technological solutionism. The main concern of machine ethicists today is still that of building tools which do not generate a moral debt to the society in which they are implemented; and not yet the project of building enlightened

Philanthropotrons. Indeed, with the increasing ubiquity of technology, it stands to reason that one should be *more* concerned with the acceptability of these machines, rather than doubting the cogency of the moral attitudes expressed by its users.

Accordingly, many of the arguments of this thesis—the place of artificial morality, the threshold of moral blindness, and even the Ethical Valence Theory itself—can be seen as attempts to capture what we might call the *moral minimum* in the design of artificial morality. This doctrine attempts to place the moral behavior of machines in the center of two opposing poles: on one hand, the amoral and prudential reasoning at work behind the use and purchase of many machines, and on the other, the best moral outcome that we are able to hope for given what we know about morality. In this way, the moral minimum is an improvement upon the status quo, but not upon the human condition.

This position seems especially reasonable in light of the *speed* at which these ethical questions rise and fall in the literature and beyond. When this thesis was first begun, the idea of machine ethics was still quite speculative, and the idea that machine ethics itself would become a premier topic on the tongues of the world's most powerful leaders almost laughable. In this sense, in the span of less than half a decade, the machine ethics community, and particularly that of autonomous vehicles, went from fighting for legitimacy to fending off media backlash. While increased public awareness concerning the ethics of technology is almost always a good thing, it does have its drawbacks: the 'pigeon holeing' of serious and delicate philosophical concepts into crude and superficial categories, the reticence of stakeholders to pronounce meaningful positive statements about ethics, lest they be hounded by the press, an increasingly generalized misunderstanding of the role of ethics in technology from which it will be difficult to return, and finally, an unfounded vision of the philosopher's character, and his ability to provide concrete, practical and perfect answers to fundamentally unanswerable problems.

All this to say that a philosopher working in the ethics of technology today is often expected to have a near clairvoyant vision of our virtual future, and to make predictions whose truth value borders on that of religious prophesies. It seems important here to recall that philosophy is not a predictive, or even a truly practical venture; nor is it meant to provide scripture for the

technological gurus of the world. At best, philosophy can point out those intricacies of the world which fall between the cracks of different disciplinary paradigms, and offer possible worlds in which some of these cracks are sealed. In this sense, this thesis has pointed out one such crack between morality in theory and morality in practice, and has attempted to provide a vision of a possible world in which this tension is resolved in an acceptable way.

References

Abbeel, P., & Ng, A. Y. (2004, July). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning* (p. 1).

Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot ethics: The ethical and social implications of robotics*, 35-52.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Adams, T. K. (2001). Future warfare and the decline of human decision-making. *Parameters*, *31*(4), 57-71.

Aldewereld, H., Álvarez-Napagao, S., Dignum, F., & Vázquez-Salceda, J. (2010, May). Making norms concrete. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1* (pp. 807-814).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, *7*(3), 149-155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251-261.

Allen, C., & Wallach, W. (2012). Moral machines: contradiction in terms or abdication of human responsibility. *Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, Cambridge (MA)*, 55-68.

Airbib, J. & Seba, T. (2017). Rethinking transportation 2020-2030: The Disruption of Transportation and the Collapse of the Internal Combustion Vehicle and Oil Industries. *RethinkX: Rethink Transportation*.

Anderson, S. L., & Anderson, M. (2015). Towards a principle-based healthcare agent. In *Machine Medical Ethics* (pp. 67-77). Springer, Cham.
Anderson, J. M., Nidhi, K., Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2014). *Autonomous vehicle technology: A guide for policymakers*. Rand Corporation.

Anderson, M., & Anderson, S. L. (2014). Toward ethical intelligent autonomous healthcare agents: A case-supported principle-based behavior paradigm. In *Proceedings of the 50th Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB-50) Symposium on Machine Ethics in the Context of Medical and Care Agents, London, UK*.

Anderson, M., & Anderson, S. L. (2010). Robot be good. *Scientific American*, *303*(4), 72-77.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *Ai Magazine*, *28*(4), 15-15.

Anderson, K., & Waxman, M. C. (2012). Law and ethics for robot soldiers. *Policy Review*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Retrieved May 10th, 2020 from: *https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*.

Applbaum, A. I. (1999). *Ethics for adversaries: The morality of roles in public and professional life*. Princeton University Press.

Arkin, R. (2015). The case for banning killer robots: counterpoint. *Communications of the ACM*, *58*(12), 46-47.

Arkin R. (2009), *Governing Lethal Behavior in Autonomous Robots*, New York, Chapman & Hall/ CRC Press.

Arkin, R. C. (2008, March). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 121-128).

Arkin, R. C., Fujita, M., Takagi, T., & Hasegawa, R. (2003). An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems*, *42*(3-4), 191-201.

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, *18*(2), 103-115.

Asaro, P. M. (2016, March). The liability problem for autonomous artificial agents. In *2016 AAAI Spring Symposium Series*.

Asaro, P. M. (2006). What should we want from a robot ethic. *International Review of Information Ethics*, *6*(12), 9-16.

Ashford, E., & Mulgan, T. (2018) Contractualism. *The Stanford Encyclopaedia of Philosophy.* Retrieved from: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=contractualism

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, F. & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Bahouth, G., Graygo, J., Digges, K., Schulman, C., & Baur, P. (2014). The benefits and tradeoffs for varied high-severity injury risk thresholds for advanced automatic crash notification systems. *Traffic injury prevention*, *15*(sup1), S134-S140.

Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.*, *31*, 889.

Bauer, W. A. (2018). Virtuous vs. utilitarian artificial moral agents. *AI & SOCIETY*, *35*(1), 263-271.

Bauer, W. A. (2020). Expanding Nallur's Landscape of Machine Implemented Ethics. *Science and Engineering Ethics*, 1-10.

Beavers, A. F. (2011). Moral Machines and the Threat of Ethical Nihilism in Lin, P., Abney, K., & Bekey, G. (eds.), *Robot ethics: The ethical and social implications of robotics*, 330-343.

Behdadi, D., & Munthe, C. (2018). Artificial Moral Agency: Philosophical Assumptions, Methodological Challenges, and Normative Solutions. *Manuscript under Consideration for Publication)*.

Beiker, S. A. (2012). Legal aspects of autonomous driving. *Santa Clara L. Rev.*, *52*, 1145.

Bekker, I., Bolland, W.E., Lang, A., Aristotle. (1877). *Aristotle's Politics: Books I, III, IV, (VII).* Longmans Green, London.

Belliotti, R. A. (1981). Negative and positive duties. *Theoria*, *47*(2), 82-92.

Bello, P., & Bridewell, W. (2017). There is no agency without attention. *AI Magazine*, *38*(4), 27-34.

Berlin, I. (2017). Two concepts of liberty. In *Liberty Reader* (pp. 33-57). Routledge.

Berreby, F., Bourgne, G., & Ganascia, J. G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for programming, artificial intelligence, and reasoning* (pp. 532-548). Springer, Berlin, Heidelberg.

Bhargava, V., & Kim, T. W. (2017). Autonomous vehicles and moral uncertainty, in Lin, P., Abney, K., & Jenkins, R., (eds) *Robot Ethics*, *2.0: From Autonomous Cars to Artificial Intelligence.*

Blackorby, Charles, Walter Bossert, and David J. Donaldson. *Population issues in social choice theory, welfare economics, and ethics*. No. 39. Cambridge University Press, 2005.

Blackorby, C., Bossert, W., & Donaldson, D. J. (1997). Critical-level utilitarianism and the population-ethics dilemma.

BMVI, (2017). Ethics Commission: Automated and Connected Driving. Federal Ministry of Transport and Digital Infrastructure, Germany. Retrieved from: https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, *20*(1), 41-58.

Borenstein, J., & Arkin, R. (2016). Robotic nudges: the ethics of engineering a more socially just human being. *Science and engineering ethics*, *22*(1), 31-46.

Bose, D., Segui-Gomez, ScD, M., & Crandall, J. R. (2011). Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk. *American journal of public health*, *101*(12), 2368-2373.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 277-284.

Bostrom, N. S. (2014). Superintelligence: Paths, Dangers, Strategies.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, *1*, 316-334.

Bradshaw, J. M., Beautement, P., Breedy, M. R., Bunch, L., Drakunov, S. V., Feltovich, P. J., ... & Lott, J. (2004). Making agents acceptable to people. In *Intelligent Technologies for Information Analysis* (pp. 361-406). Springer, Berlin, Heidelberg.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of" autonomous systems". *IEEE Intelligent Systems*, *28*(3), 54-61.

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. In *Springer handbook of robotics* (pp. 1935-1972). Springer, Cham.

Bringsjord, S., Ghosh, R., & Payne-Joyce, J. (2016). Deontic counter identicals. *Agents (EDIA)*, *2016*, 40-45.

Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. *Robot ethics: The ethical and social implications of robotics*, 85-108.

Bringsjord, S. (2008). Ethical robots: the future can heed us. *Ai & Society*, *22*(4), 539-550.

Brink, D. O. (1986). Utilitarian morality and the personal point of view. *The Journal of Philosophy*, *83*(8), 417-438.

Brooks, R. A. (2003). *Flesh and machines: How robots will change us*. Vintage.

Brooks, R. A. (1991). New approaches to robotics. *Science*, *253*(5025), 1227-1232.

Broome, J. (2017). *Weighing goods: Equality, uncertainty and time*. John Wiley & Sons.

Broome, J. (2004). Weighing lives. *OUP Catalogue*.

Brustoloni, J. C. (1991). *Autonomous agents: Characterization and requirements*. School of Computer Science, Carnegie Mellon University.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15-26.

Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 63-74.

Bryson, J. J., & Kime, P. P. (2011, June). Just an artifact: Why machines are perceived as moral agents. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Buehler, M., Iagnemma, K., & Singh, S. (Eds.). (2009). *The DARPA urban challenge: autonomous vehicles in city traffic* (Vol. 56). springer.

Bunge, M. (1977). Towards a technoethics. *The Monist*, *60*(1), 96-107.

Casey, B. (2016). Amoral machines, or: How roboticists can learn to stop worrying and love the law. *Nw. UL Rev.*, *111*, 1347.

Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. *Communications of the ACM*, *59*(5), 70-77.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623):20.

Cellan-Jones, R. (2014, December 2). Stephen Hawking Warns Artificial Intelligence Could End Mankind. *BBC News*. Retrieved from: https://www. technology-30290540#:~:text=Prof%20Stephen%20Hawking%2C%20one%20of,end%20of%20the%20hu

man%20race%22&text=But%20others%20are%20less%20gloomy%20about%20AI's%20prospects.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy & Technology*, *28*(1), 125-137.

Chatila, R., Firth-Butterflied, K., Havens, J. C., & Karachalios, K. (2017). The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. *IEEE Robotics & Automation Magazine*, *24*(1).

Chaudhuri, S., & Vardi, M. Y. (2014). Reasoning about machine ethics. Presentation available at: https://popl-obt-2014.cs.brown.edu/papers/ethics.pdf.

Chauvier, S. (2013). *Éthique sans visage : le problème des effets externes*. Librairie philosophique J. Vrin.

Chopra, A. K., & Singh, M. P. (2018, December). Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 48-53).

Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. *Advances in human performance and cognitive engineering research*, *2*, 1-12.

Cicero, M. T. (1913). De officiis, trans. Walter Miller. *London: William Heinemann Ltd*, *27*, 3-64.

Clamann, M., Aubert, M., & Cummings, M. L. (2017). *Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles* (No. 17-02119).

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748-757.

Coeckelbergh, M. (2011). Humans, animals, and robots: A phenomenological approach to human-robot relations. *International Journal of Social Robotics*, *3*(2), 197-204.

Coeckelbergh, M. (2012). Can we trust robots?. *Ethics and information technology*, *14*(1), 53-60.

Coeckelbergh, M. (2009). Personal robots, appearance, and human good: A methodological reflection on roboethics. *International Journal of Social Robotics*, *1*(3), 217-221.

Cohen, J., & Sabel, C. (2006). Extra rempublicam nulla justitia?. *Philosophy & public affairs*, *34*(2), 147-175.

Condliffe, J. (2016, November 3). Humans Will Bully Mild-Mannered Autonomous Cars. *MIT Technology Review*. Retrieved from: https://www.technologyreview.com/2016/11/03/156270/humans-will-bully-mild-mannered-autonomous-cars/

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017, February). Moral Decision Making Frameworks for Artificial Intelligence. In *AAAI Workshops*.

Conitzer, V., Brill, M., & Freeman, R. (2015, May). Crowdsourcing societal tradeoffs. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1213-1217).

Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365-378.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311-313.

Crocetta, G., Piantini, S., Pierini, M., & Simms, C. (2015). The influence of vehicle front-end design on pedestrian ground impact. *Accident Analysis & Prevention*, *79*, 56-69.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299-309.

Dancy, J. (2004). *Ethics without principles*. Oxford University Press on Demand.

Davies, C. (2019, April 10). Tesla Autopilot Safety Report. *Slashgear*. Retrieved from: https://www.slashgear.com/tesla-autopilot-safety-report-q1-2019-autonomous-progress-10572659/.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Dietrich, F., & List, C. (2017). What matters and how it matters: a choice-theoretic representation of moral theories. *Philosophical Review*, *126*(4), 421-479.

Dizikes, P. (2018, October 24). How Should Autonomous Vehicles Be Programmed? *MIT News*. Retrieved from: http://news.mit.edu/2018/how-autonomous-vehicles-programmed-1024

Dennett, D. C. (1998a). When HAL Kills, Who's to Blame? In Stork, D. G. (Ed.). *HAL's Legacy: 2001's Computer as Dream and Reality*. MIT Press.

Dennett, D. C. (1998b). *Brainchildren: Essays on designing minds*. MIT Press.

Dennis, L., & Fisher, M. (2018). Practical challenges in explicit ethical machine reasoning. *arXiv preprint arXiv:1801.01422*.

De Sio, F. S. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice*, *20*(2), 411-429.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809-828.

Dietrich, E. (2001). Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 323-328.

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing.

Docherty, B. L. (2015). *Mind the gap: The lack of accountability for killer robots*. Human Rights Watch.

Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., & Perrin, J. (2016, August). Ethics in the Design of Automated Vehicles: The AVEthics project. In *EDIA@ ECAI* (pp. 10-13).

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1).

Drozdek, A. (1992). Moral dimension of man and artificial intelligence. *AI & society*, *6*(3), 271-280.

Dworkin, R. (2013). *Taking rights seriously*. A&C Black.

Eifling, S. (2016, May 19). Google Patented a Sticky Car Hood to Protect Pedestrians Who Get Hit. *Popular Mechanics*. Available at:https://www.popularmechanics.com/technology/a20953/google-patent-sticky-car-hood/

Eilam, E. (2005). Reversing. *Secrets of Reverse Engineering, Indianapolis*.

Epstein, S., & Pacini, R. (1999). Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. *Dual-process theories in social psychology*, 462-482.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403-418.

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project. *Science and Engineering Ethics*, 1-28.

Evans, N. G. (2019, July). Ethical Algorithms in Autonomous Vehicles: Reflections on a Workshop. In *Automated Vehicles Symposium* (pp. 161-169). Springer, Cham.

Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382-389.

Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., ... & Petersen, S. (2018). De novo structure prediction with deeplearning based scoring. *Annu Rev Biochem*, *77*(363-382), 6.

Ezenkwu, C. P., & Starkey, A. (2019, July). Machine autonomy: Definition, approaches, challenges and research gaps. In *Intelligent Computing-Proceedings of the Computing Conference* (pp. 335-358). Springer, Cham.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77, 167-181.*

Feinberg, J. (1987). *Harm to others* (Vol. 1). Oxford University Press on Demand.

Firenze, A. (2019). Lacking what? On the Welt-Umwelt dichotomy in Heidegger and Gehlen. *Enrahonar. An International Journal of Theoretical and Practical Reason*, *63*, 39-53.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., ... & Dragan, A. D. (2020). Pragmatic-pedagogic value alignment. In *Robotics Research* (pp. 49-57). Springer, Cham.

Fitts, P. M. (1951). Human engineering for an effective air-navigation and traffic-control system.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American journal of public health*, *107*(4), 532-537.

Floridi, L. (2008). Information Ethics: Its Nature and Scope, in Van Den Hoven, J., & Weckert, J. (Eds.). *Information technology and moral philosophy*. Cambridge University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14*(3), 349-379.

Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3*(1), 55-66.

Foot, P. (1967). The problem of abortion and the doctrine of double effect.

Froese, T., Virgo, N., & Izquierdo, E. (2007, September). Autonomy: a review and a reappraisal. In *European Conference on Artificial Life* (pp. 455-464). Springer, Berlin, Heidelberg.

Frank, D. A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific reports*, *9*(1), 1-19.

Franklin, S., & Graesser, A. (1996, August). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International Workshop on Agent Theories, Architectures, and Languages* (pp. 21-35). Springer, Berlin, Heidelberg.

Freitas, J. D., Anthony, S. E., & Alvarez, G. (2019). Doubting driverless dilemmas. *PsyArXiv Preprints DOI*, *10*.

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. *The handbook of information and computer ethics*, 69-101.

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *arXiv preprint arXiv:2001.09768*.

Ganascia, J. G. (2007). Ethical system formalization using non-monotonic logics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29, No. 29).

Gao, P., Kaas, H. W., Mohr, D. & Wee, D. (2016). *Automotive revolution-perspective towards 2030: How the convergence of disruptive technology-driven trends could transform the auto industry.* Advanced Industries, McKinsey & Company.

Gerdes, A., & Øhrstrøm, P. (2013, June). Preliminary reflections on a moral Turing test. In *Proceedings of ETHICOMP* (pp. 167-174).

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In *Autonomes fahren* (pp. 87-102). Springer Vieweg, Berlin, Heidelberg.

Gert, B. (2004). *Common morality: Deciding what to do*. Oxford University Press.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

Gips, J. (1995, October). Towards the ethical robot. In *Android epistemology* (pp. 243-252). MIT Press.

Glancy, D. J. (2012). Privacy in autonomous vehicles. *Santa Clara L. Rev.*, *52*, 1171.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: in favor of a mandatory ethics setting. *Science and engineering ethics*, *23*(3), 681-700.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In *Road vehicle automation* (pp. 93-102). Springer, Cham.

Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810-821.

Goodall, N. J. (2017). From trolleys to risk: Models for ethical autonomous driving. *American Journal of Public Health, 107(4), 490-502.*

Goodall, N. (2019). More than Trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles. *Transfers*, *9*(2), 45-58.

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. *arXiv preprint arXiv:1703.08922*.

de Graaf, M. M., Ben Allouch, S., & van Dijk, J. A. (2019). Why would I use this in my home? A model of domestic social robot acceptance. *Human–Computer Interaction*, *34*(2), 115-173.

de Graaf, M. M. (2016). An ethical evaluation of human–robot relationships. *International journal of social robotics*, *8*(4), 589-598.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. C. (2016, February). Embedding Ethical Principles in Collective Decision Support Systems. In *Aaai* (Vol. 16, pp. 4147-4151).

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144-1154.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Griggs, T. (2018, March 20). "How a self-driving uber killed a pedestrian in Arizona. *The New York Times.* Retrieved from: https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html?mtrref=www.google.com&gwh=02AF916CCA74980774429EDAA158B560&gwt=pay&assetType=REGIWALL

Grinbaum, A. (2018). Chance as a value for artificial intelligence. *Journal of Responsible Innovation, 5*(3), 353-360.

Grinbaum, A. (2019). *Les robots et le mal*. Desclée De Brouwer.

Gunkel, D. J. (2017). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, 1-14.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Gurney, J. K. (2015). Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Alb. L. Rev.*, *79*, 183.

Hall, J. S. (2009). *Beyond AI: Creating the conscience of the machine*. Prometheus books.

Hao, K. (2020, March 6). A Hybrid AI Model Lets It Reason About the World's Physics Like a Child. *MIT Technology Review.* Retrieved from: https://www.technologyreview.com/2020/03/06/905479/ai-neuro-symbolic-system-reasons-like-child-deepmind-ibm-mit/

Heidegger, M. (1997). *The Question Concerning Technology, and Other Essays.* New York, Harper & Row, 1977. p.3.

Heidegger, M. (1995). *The fundamental concepts of metaphysics: World, finitude, solitude*. Indiana University Press.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology*, *11*(1), 19-29.

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*(3), 669-684.

Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books.

Hsu, J. (2014, December 31). Google's Car is the Face of Future Robots. *Discover Magazine*. Retrieved from: *https://www.discovermagazine.com/technology/googles-car-is-the-face-of-future-robots*

Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In *Philosophy and computing* (pp. 121-159). Springer, Cham.

Hübner, D., & White, L. (2018). Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice*, *21*(3), 685-698.

IEEE Global Initiative. (2016). Ethically aligned design. *IEEE Standards V1*.

Irrgang, B. (2006). Ethical acts in robotics. *Ubiquity*, *2006*(September), 2-16.

Jaques, A. E. (2019). Why the moral machine is a monster. *University of Miami School of Law*, *10*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Johannsen, G. (2009). Human-machine interaction. *Control Systems, Robotics and Automation*, *21*, 132-62.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and information technology*, *8*(4), 195-204.

Johnson, D. G. (1985/ 2001). *Computer ethics*, 3rd ed. Englewood Cliffs (NJ).

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines*, *27*(4), 575-590.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology*, *10*(2-3), 123-133.

Johnson, D., & Powers, T. M. (2008). Computers as surrogate agents. *Information technology and moral philosophy*, 251-269.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and information technology*, *7*(2), 99-107.

Jonas, H. (1979). Toward a philosophy of technology. *Hastings Center Report*, 34-43.

Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., & Corben, B. (2016). Exploration of vehicle impact speed–injury severity relationships for application in safer road design. *Transportation research procedia*, *14*, 4247-4256.

Kagan, S. (1994). The argument from liberty. *In Harm's Way: Essays in Honor of Joel Feinberg*, 16-41.

Kagan, S. (1992). The structure of normative ethics. *Philosophical perspectives, 6*, 223-242.

Kagan, S. (1989). *The limits of morality*. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kamm, F. M. (2016). *The Trolley Problem Mysteries*. New York: Oxford University Press.

Kant, I. (2013). *An answer to the question:'What is enlightenment?'*. Penguin UK.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keeling, G. (2020a). Why trolley problems matter for the ethics of automated vehicles. *Science and engineering ethics*, *26*(1), 293-307.

Keeling, G. (2020b). *The Ethics of Automated Vehicles* (Doctoral dissertation, University of Bristol).

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & de Sio, F. S. (2019, July). Four perspectives on what matters for the ethics of automated vehicles. In *Automated Vehicles Symposium* (pp. 49-60). Springer, Cham.

Keeling, G. (2018). Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions. *Ethical Theory and Moral Practice*, *21*(2), 413-427.

Keeling, G. (2017). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In *3rd Conference on" Philosophy and Theory of Artificial Intelligence* (pp. 259-272). Springer, Cham.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis & Prevention*, *40*(5), 1695-1702.

Kiverstein, J. (2007). Could a robot have a subjective point of view?. *Journal of Consciousness Studies*, *14*(7), 127-139.

Klincewicz, M. (2017). Callenges to Engineering Moral Reasoners: Time and Context. In Lin, P., Jenkins, R. & Abney, K. (Eds.). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 244-257).

Knight, W. (2019, April 8). Two Rival AI Approaches Combine To Let Machines Learn About the World Like A Child. *MIT Technology Review*. Retrieved from: https://www.technologyreview.com/2019/04/08/103223/two-rival-ai-approaches-combine-to-let-machines-learn-about-the-world-like-a-child/

Kröyer, H. R. (2015). Is 30 km/ha 'safe'speed? Injury severity of pedestrians struck by a vehicle and the relation to travel speed and age. *IATSS research*, *39*(1), 42-50.

Kuipers, B. (2018). How can we trust a robot?. *Communications of the ACM*, *61*(3), 86-95.

LaGrandeur, K. (2013). *Androids and Intelligent Networks in Early Modern Literature and Culture: Artificial Slaves.* Routledge.

Lampe, A., & Chatila, R. (2006, May). Performance measure for the evaluation of mobile robot autonomy. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* (pp. 4057-4062). IEEE.

Langheinrich, M. (2001, September). Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on Ubiquitous Computing* (pp. 273-291). Springer, Berlin, Heidelberg.

L.A. Times, Associated Press (2019, September 3). *Tesla car was on Autopilot when it hit a Culver City firetruck, NTSB finds.* L.A. Times, retrieved 10-05-2020 from https://www.latimes.com/business/story/2019-09-03/tesla-was-on-autopilot-when-it-hit-culver-city-fire-truck-ntsb-finds.

Leben, D. (2018). *Ethics for robots: How to design a moral algorithm*. Routledge.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19*(2), 107-115.

Leenes, R., & Lucivero, F. (2014). Laws on robots, laws by robots, laws in robots: regulating robot behavior by design. *Law, Innovation and Technology*, *6*(2), 193-220.

Levy, D. (2009a). *Love and sex with robots: The evolution of human-robot relationships*. New York.

Levy, D. (2009b). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics*, *1*(3), 209-216.

Liu, J., Hainen, A., Li, X., Nie, Q., & Nambisan, S. (2019). Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach. *Accident Analysis & Prevention*, *132*, 105272.

Lin, P. (2017). Robot Cars and Fake Ethical Dilemmas. *Forbes Magazine.* Retrieved from: https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/#53bd0e2213a2

Lin, P. (2016). Why ethics matters for autonomous cars. In *Autonomous driving* (pp. 69-85). Springer, Berlin, Heidelberg.

Lin, P. (2014a). Here's a terrible idea: robot cars with adjustable ethics settings. *Wired. com. Available online at: http://www. wired. com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings*.

Lin, P. (2014b, May 6). "The Robot Car of Tomorrow May Just Be Programmed to Hit You". *Wired*. Retrieved from: *http://wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/*.

Lin, P., Mehlman, M., Abney, K., & Galliott, J. (2014). Super Soldiers (Part 1): What is Military Human Enhancement?. In *Global issues and ethical considerations in human enhancement technologies* (pp. 119-138). IGI Global.

Lin, P. (2013). The ethics of autonomous cars. *The Atlantic*, *8*.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. California Polytechnic State Univ San Luis Obispo.

Litman, T. (2020). *Autonomous Vehicle Implementation Predictions: Implications for Transport Planning.* Victori Transport Policy Institute.

Litman, T. (2017). *Autonomous vehicle implementation predictions*. Victoria, Canada: Victoria Transport Policy Institute.

Loh, W., & Loh, J. (2017). Autonomy and responsibility in hybrid systems—the example of autonomous cars. In: Lin, P., Abney, K., & Jenkins, R. (*eds*) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, New York, pp. 35-50.

Loh, W., & Misselhorn, C. (2019). Autonomous Driving and Perverse Incentives. *Philosophy & Technology*, *32*(4), 575-590.

MacAskill, W. (2016). Normative uncertainty as a voting problem. *Mind*, *125*(500), 967-1004.

MacKenzie, E. J., Shapiro, S., & Eastham, J. N. (1985). The Abbreviated Injury Scale and Injury Severity Score: Levels of Inter and Intrarater Reliability. *Medical Care*, 823-835.

Maes, P. (1995). Artificial life meets entertainment: lifelike autonomous agents. *Communications of the ACM*, *38*(11), 108-114.

Mahadevan, K., Somanath, S., & Sharlin, E. (2018, April). Communicating awareness and intent in autonomous vehicle-pedestrian interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Malczyk, A., Müller, G., & Gehlert, T. (2012). *The Increasing Role of SUVs in Crash Involvement in Germany*. Ireland: IRCOBI.

Malle, B. F., & Scheutz, M. (2018). Learning how to behave. Moral competence for social robots. *Springer, Wiesbaden, Germany*, 1-24.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125-132).

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 117-124). IEEE.

Maner, W. (1996). Unique Ethical Problems in Information Technology. In Bynum, T.W. & Rogerson, S. (eds.), *Global Information Ethics*. Special Issue of *Science and Engineering Ethics*, 2(2): 137-154.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Martin, J. L., & Wu, D. (2018). Pedestrian Fatality and Impact Speed Squared: Cloglog Modelling from French National Data. *Traffic Injury Prevention*, 19(1), 94-101.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12-12.

McDermott, D. (2008, February). Why ethics is a high hurdle for AI. In *North american conference on computers and philosophy*.

McFarland, T., & McCormack, T. (2014). Mind the Gap: Can Developers of Autonomous Weapons Systems be Liable for War Crimes?. *International Law Studies*, *90*(1), 2.

McLaren, B. M. (2003). Extensionally defining principles and cases in ethics: An AI model. *Human-Computer Interaction Institute*, 130.

Meder, B., Fleischhut, N., Krumnau, N. C., & Waldmann, M. R. (2019). How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk analysis*, *39*(2), 295-314.

Miceli, M., & Castelfranchi, C. (1989). A cognitive approach to values. *Journal for the Theory of Social Behaviour*, *19*(2), 169-193.

Millar, J., Lin, P., Abney, K., & Bekey, G. A. (2017). *Ethics settings for autonomous vehicles* (pp. 20-34). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence.* MIT Press.

Millar, J., & Kerr, I. (2016). Delegation, relinquishment, and responsibility: The prospect of expert robots. In *Robot Law*. Edward Elgar Publishing.

Millar, J. (2015). Technology as moral proxy: Autonomy and paternalism by design. *IEEE Technology and Society Magazine*, *34*(2), 47-55.

Millar, J. (2014a). Proxy Prudence: Rethinking Models of Responsibility for Semi-Autonomous Robots. *Available at SSRN 2442273*.

Millar, J. (2014b, September 2). You should have a say in your robot car's code of ethics. *WIRED. com*. Retrieved from: http://www.wired.com/2014/09/set-the-ethics-robot-car/.

Millard-Ball, A. (2018). Pedestrians, autonomous vehicles, and cities. *Journal of planning education and research*, *38*(1), 6-12.

Mizuno, K., & Kajzer, J. (1999). Compatibility problems in frontal, side, single car collisions and car-to-pedestrian accidents in Japan. *Accident Analysis & Prevention*, *31*(4), 381-391.

Mladenovic, M. & McPherson, T. (2016). Engineering Social Justice Into Traffic Control for Self-Driving Vehicles. *Science and Engineering Ethics* 22:1131-49.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, *21*(4), 18-21.

Moor, J. H. (2005). Why we need better ethics for emerging technologies. *Ethics and information technology*, *7*(3), 111-119.

Moor, J. H. (1979). Are there decisions computers should never make. *Nature and System*, *1*(4), 217-229.

De Moura, N., Chatila, R., Evans, K., Chauvier, S., & Dogan, E. (2020). Ethical Decision-Making for Autonomous Vehicles. *IEEE Intelligent Vehicles Symposium (iv).*

Nadeau, J. E. (2006). Only Androids Can be Ethical. In Ford, K., Glymour, C., & Hayes, P.J., (eds), *Thinking About Android Epistemology* (pp. 241-248). Cambridge: MIT Press.

Nagel, T. (2012). *Mortal questions*. Cambridge University Press.

Nallur, V. (2020). Landscape of machine implemented ethics. *Science and engineering ethics*, 1-19.

Naughton, K. (2015). Humans Are slamming into driverless cars and exposing a key flaw: Bloomberg business. Retrieved February 23, 2016, from http://www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-

driverless-cars-and-exposing-a-key-flaw?utm_content=buffer16029&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer.

Newell, A. (1982). The knowledge level. *Artificial intelligence*, *18*(1), 87-127.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Icml* (Vol. 1, p. 2).

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and engineering ethics*, *2*(1), 25-42.

Nolfi, S., Floreano, D., & Floreano, D. D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT press.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51-62.

Noothigattu, R., Gaikwad, S. N. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2017). A voting-based system for ethical decision making. *arXiv preprint arXiv:1709.06692*.

Norman, D. A. (1991). Cognitive artifacts, Designing interaction: psychology at the human-computer interface.

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, *24*(4), 1201-1219.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507.

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: a roadmap, II. *Philosophy Compass*, *13*(7), e12506.

Nyholm, S., & Smids, J. (2018). Automated cars meet human drivers: Responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*, 1-10.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem?. *Ethical theory and moral practice*, *19*(5), 1275-1289.

Ogien, R. (2007). L'éthique aujourd'hui. *Maximalistes et minimalistes*, 144-152.

Oh, S. J., Schiele, B., & Fritz, M. (2019). Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 121-144). Springer, Cham.

Ottawa Citizen, The. (1983, August 11). *$10 Million Awarded to Family of Plant Worker Killed By Robot, 14.* Retrieved from: https://news.google.com/newspapersid=7KMyAAAAIBAJ&sjid=Bu8FAAAAIBAJ&pg=3301,87702&dq=flat-rock+williams+robot&hl=en.

Parfit, D. (2011). On What Matters (Vol. 2). Oxford: Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. OUP Oxford.

Pereira, L. M., & Saptawijaya, A. (2007, December). Modelling morality with prospective logic. In *Portuguese Conference on Artificial Intelligence* (pp. 99-111). Springer, Berlin, Heidelberg.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Peters, J., Kober, J., Mülling, K., Krämer, O., & Neumann, G. (2013, September). Towards robot skill learning: From simple skills to table tennis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 627-631). Springer, Berlin, Heidelberg.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, *117*(2), 171-201.

Pickering, J. E., Ashman, P., Gilbert, A., Petrovic, D., Warwick, K., & Burnham, K. J. (2018, September). Model-to-Decision Approach for Autonomous Vehicle Convoy Collision Ethics. In *2018 UKACC 12th International Conference on Control (CONTROL)* (pp. 301-308). IEEE.

Pitt, J. C. (2011). *Doing philosophy of technology: essays in a pragmatist spirit* (Vol. 3). Springer Science & Business Media.

Podschwadek, F. (2017). Do androids dream of normative endorsement? On the fallibility of artificial moral agents. *Artificial Intelligence and Law*, *25*(3), 325-339.

Poulsen, A., Anderson, M., Anderson, S. L., Byford, B., Fossa, F., Neely, E. L., ... & Winfield, A. (2019). Responses to a Critique of Artificial Moral Agents. *arXiv preprint arXiv:1903.07021*.

Powers, T. M. (2013). Prospects for a Smithian machine. *Proceedings of the international association for computing and philosophy, College Park, Maryland*.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems*, *21*(4), 46-51.

Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, *25*(3), 341-363.

Prasad, M. (2018). Social choice and the value alignment problem. In *Artificial intelligence safety and security* (pp. 291-314). Chapman and Hall/CRC.

Prescott, T. J. (2017). Robots are not just tools. *Connection Science*, *29*(2), 142-149.

Price, W. N. (2017). Medical Malpractice and Black-Box Medicine. *Big Data, Health Law, and Bioethics (Cambridge University Press, 2018)*.

Prichard, H. A. (2002/1932). Duty and Ignorance of Fact. In Moral Writings (pp. 84-101). Oxford: Clarendon Press.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, *18*(4), 851-872.

Qian, X., Navarro, I., de La Fortelle, A. & Moutarde, F. (2016) Motion planning for urban autonomous driving using Bézier curves and MPC. In *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)* (pp. 826–833).

Rawls, J. (2009). *A theory of justice*. Harvard university press.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Reeves, B., & Nass, C. (1996). How people treat computers, television, and new media like real people and places. CSLI, New York.

Rizaldi, A., Keinholz, J., Huber, M., Feldle, J., Immler, F., Althoff, M. & Nipkow, T. (2017) Formalising and monitoring traffic rules for autonomous vehicles in Isabelle/HOL. In *International conference on integrated formal methods* (pp. 50–66).

Rosen, E., Stigson, H., & Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis & Prevention*, *43*(1), 25-33.

Rouchitsas, A., & Alm, H. (2019). External Human–Machine Interfaces for Autonomous Vehicle-to-Pedestrian Communication: A Review of Empirical Work. *Frontiers in psychology*, *10*.

Roy, A. (2016, October 19). Autonomous Cars Don't Have a 'Trolley Problem' Problem. *The Drive*. Available at: http://www.thedrive.com/tech/5620/autonomous-cars-dont-have-a-trolleyproblem-problem.

Russel, S., & Norvig, P. (2013). *Artificial intelligence: a modern approach*. Pearson Education Limited.

Russel, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* Pearson Education Limited.

SAE (Society of Automative Engineers). (2016, September). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.* Retrieved from: *https://www.sae.org/standards/content/j3016_201609/.*

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, October). Towards safe and trustworthy social robots: ethical challenges and practical issues. In *International conference on social robotics* (pp. 584-593). Springer, Cham.

Salimans, T., Ho, J., Chen, X., Sidor, S., & Sutskever, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*.

Sandberg, A., & Bradshaw, H. G. (2013). Autonomous vehicles, moral agency and moral proxyhood. In *Beyond AI conference proceedings. Springer*.

Scanlon, T. (1998). *What We Owe to Each Other.* Harvard University Press.

Scanlon, T. (2008). *Moral dimensions: Permissibility, meaning, blame*. Cambridge, MA: Belknap.

Scanlon, T. M. (1982). In Sen, A., Williams, B., & Williams, B. A. O. (Eds.). *Utilitarianism and beyond*. Cambridge University Press.

Scheutz, M. (2017). The case for explicit ethical agents. *AI Magazine*, *38*(4), 57-64.

Scheutz, M. (2016). The need for moral competency in autonomous agent architectures. In *Fundamental issues of artificial intelligence* (pp. 517-527). Springer, Cham.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694-696.

Sharkey, N. E. (2012). The evitability of autonomous robot warfare. *International Review of the Red Cross*, *94*(886), 787-799.

Siciliano, B., & Khatib, O. (Eds.). (2016). *Springer handbook of robotics*. Springer.

Sidgwick, H. (2019). *The methods of ethics*. Good Press.

Sigaud, O., & Buffet, O. (2013). Markov decision processes in artificial intelligence. Hoboken: Wiley.

Simms, C. K., & Wood, D. P. (2006). Pedestrian risk from cars and sport utility vehicles-a comparative analytical study. *Proceedings of the Institution of Mechanical Engineers Part D: Journal of Automobile Engineering*, *220*(8), 1085–1100.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & public affairs*, 229-243.

Slote, M. (1985). *Common-sense Morality and Consequentialism.* London: Rout-ledge & Kegan Paul.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies*, *80*, 206-215.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies*, *80*, 206-215.

Spiekermann, S. (2015). *Ethical IT innovation: A value-based system design approach*. CRC Press.

Stahl, B. C. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines*, *14*(1), 67-83.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, *86*, 152-161.

Stern, R. (2018, March 9). Tempe Police: Uber Self-Driving Car Didn't Brake 'Significantly' Before Killing Pedestrian. *Phoenix New Times.* Retrieved from: https://www.phoenixnewtimes.com/news/cops-uber-self-driving-car-didnt-brake-much-in-arizona-pedestrian-death-10247629

Stewart, J. (2018, October 18). Why People Keep Rear-Ending Self-Driving Cars. *Wired Magazine.* Retrieved from: https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics/

Sullins, J. P. (2014). Ethical trust in the context of robot assisted surgery. In *AISB 2014-50th Annual Convention of the AISB.*

Sullins, J. P. (2009). Artificial moral agency in technoethics. In *Handbook of research on technoethics* (pp. 205-221). IGI Global.

Sullins, J. P. (2006). When is a robot a moral agent. *Machine ethics*, 151-160.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Svoboda, E. (2019). Your robot surgeon will see you now. *Nature Outlook.* Retrieved 11-05-20 from https://www.nature.com/articles/d41586-019-02874-0.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

Talbot, B., Jenkins, R., & Purves, D. (2017). When robots should do the wrong thing. *Robot Ethics*, *2*, 258-273.

Tavani, H. T. (2015). Levels of trust in the context of machine ethics. *Philosophy & Technology*, *28*(1), 75-90.

Tavani, H. T. (2002). The uniqueness debate in computer ethics: What exactly is at issue, and why does it matter?. *Ethics and Information Technology*, *4*(1), 37-54.
Taylor, M. (2017, October 7). Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians. *Car and Driver*. Available at: https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/.

Theodorou, A., Bandt-Law, B., & Bryson, J. J. (2019, August). The Sustainability Game: AI Technology as an Intervention for Public Understanding of Cooperative Investment. In *2019 IEEE Conference on Games (CoG)* (pp. 1-4). IEEE.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395-1415.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in Machine Ethics: A Survey. *arXiv preprint arXiv:2001.07573*.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, *19*(3), 421.

Torrance, S. (2011). Machine ethics and the idea of a more-than-human moral world. *Machine ethics*, 115-137.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *Ai &Society*, *22*(4), 495-521.

Turing, A. M. (2004). Computing machinery and intelligence (1950). *The Essential Turing: The Ideas that Gave Birth to the Computer Age. Ed. B. Jack Copeland. Oxford: Oxford UP*, 433-64.

Turkle, S. (2017). *Alone together: Why we expect more from technology and less from each other*. Hachette UK.

Turkle, S. (2005). *The second self: Computers and the human spirit*. Mit Press.

United Press International. (1981, December 8). *Robot Kills Man*. Retrieved 10-5-2020 from https://www.upi.com/Archives/1981/12/08/Robot-kills-man/2127376635600/

Van Den Hoven, J., & Weckert, J. (Eds.). (2008). *Information technology and moral philosophy*. Cambridge University Press.

Van den Hoven, J., Lokhorst, G. J., & Van de Poel, I. (2012). Engineering and the problem of moral overload. *Science and engineering ethics*, *18*(1), 143-155.

Van de Poel, I. (2016). A coherentist view on the relation between social acceptance and moral acceptability of technology. In *Philosophy of technology after the empirical turn* (pp. 177-193). Springer, Cham.

Vanderelst, D., & Winfield, A. (2018, December). The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 317-322).

Vasquez, D., Okal, B., & Arras, K. O. (2014, September). Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1341-1346). IEEE.

Verdiesen, I., Dignum, V., & Hoven, J. V. D. (2018). Measuring moral acceptability in e-deliberation: A practical application of ethics by participation. *ACM Transactions on Internet Technology (TOIT), 18*(4), 1-20.

Veruggio, G. (2005). The Birth of Roboethics. *ICRA 2005, IEEE International Conference on Robotics and Automation.*

Von Uexkull, J. (1957). A stroll through the worlds of animals and men, Instinctive Behavior. *Int. Univ. Press, New York*, 5-80.

Voorhoeve, A. (2014). How should we aggregate competing claims?. *Ethics*, *125*(1), 64-87.

Wagter, H. (2016). “Naughty Software”. Presentation at Ethics: Responsible Driving Automation, at Connekt, Delft

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015, September). Autonomous driving: investigating the feasibility of car-driver handover assistance. In *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications* (pp. 11-18).

Wallach, W., & Allen, C. (2013). Framing robot arms control. *Ethics and information technology*, *15*(2), 125-135.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wallach, W. (2015). *A dangerous master: How to keep technology from slipping beyond our control*. Basic Books.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227-248.

Weijermars, W., Bos, N., Schoeters, A., Meunier, J. C., Nuyttens, N., Dupont, E., et al. (2018). Serious road traffic injuries in europe, lessons from the eu research project safetycube. *Transportation Research Record*, *2672*(32), 1–9.

White, J. (2014). Models of moral cognition. In *Model-Based Reasoning in Science and Technology* (pp. 363-391). Springer, Berlin, Heidelberg.

Wilks, Y. (2010). Close engagements with artificial companions: Key social, psychological, ethical and design issues (Vol. 8). Amsterdam: John Benjamins Publishing.

Williams, B. (2012). *Morality: An introduction to ethics*. Cambridge University Press.

Williams, B. (2011). *Ethics and the Limits of Philosophy*. Taylor & Francis.

Williams, B. (1981). *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Williams, B. (1976). *Problems of the self: philosophical papers 1956-1972*. Cambridge University Press.

Winfield, A. F., & Jirotka, M. (2017, July). The case for an ethical black box. In *Annual Conference Towards Autonomous Robotic Systems* (pp. 262-273). Springer, Cham.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE*, *107*(3), 509-517.

de Winter, J. C., & Dodou, D. (2014). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, *16*(1), 1-11.

Wolf, S. (1997). Moral Saints. In Crisp, R., & Slote, M. (Eds.), *Virtue Ethics.* Oxford University Press.

Wooldridge, M., & Jennings, N. R. (1994, August). Agent theories, architectures, and languages: a survey. In *International Workshop on Agent Theories, Architectures, and Languages* (pp. 1-39). Springer, Berlin, Heidelberg.

World Commission on the Ethics of Scientific Knowledge and Technology (COMEST). 2017. *Report of COMEST on Robotics Ethics.* UNESDOC Digital Library. Accessible at: https://unesdoc.unesco.org/ark:/48223/pf0000253952.

van Wynsberghe, A. (2020, August). Designing Robots for Reciprocity. Keynote Presentation at RoboPhilosophy 2020. Available at: https://youtu.be/xoTq0-XFTy4

van Wynsberghe, A. (2016). Service robots, care ethics, and design. *Ethics and information technology*, *18*(4), 311-321.

van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and engineering ethics*, *19*(2), 407-433.

van Wynsberghe, A. L. (2012). Designing robots with care: Creating an ethical framework for the future design and implementation of care robots.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, *25*(3), 719-735.

Yudkowsky, E. (2016). The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.

L'Implémentation des processus de décision éthiques au sein des systèmes autonomes : le cas du véhicule autonome

Résumé

Les problèmes éthiques liés à l'émergence de nouvelles formes d'intelligence artificielle (IA) sont aujourd'hui l'objet d'intenses débats, tant académiques que publics. Une part importante des inquiétudes concerne l'IA embarquée dans des agents artificiels autonomes : comment s'assurer que les décisions prises par des agents artificiels comme des voitures autonomes ne *nuisent* pas aux êtres humains présents dans leur environnement ? Cette question a conduit à envisager que des agents artificiels autonomes socialement acceptables prennent la forme d'*agents moraux artificiels* dont la prise de décision serait contrainte par une *moralité artificielle* : un système de principes normatifs implémenté dans le processus de raisonnement et de décision de la machine. A ce jour, la forme que prend cette idée de moralité artificielle relève principalement de deux approches différentes : une forme maximaliste, qui prône l'implémentation stricte de théories morales préexistantes comme la déontologie kantienne ou l'utilitarisme dans le module décisionnel des agents artificiels ; ou bien une forme minimaliste, qui applique les techniques de l'IA stochastique à l'analyse et à l'agrégation de données portant sur les préférences morales d'une population, afin d'en tirer des principes généraux mobilisés ensuite dans la prise de décision des machines. Prises individuellement, aucune des deux approches n'arrivent à concilier les contraintes morales imposées aux agents artificiels autonomes avec les conditions de leur acceptabilité publique. Nous proposons dès lors une approche alternative, la *théorie des valences éthiques*, qui s'efforce d'accommoder cette double exigence, et nous l'appliquons au cas du véhicule autonome.

Mots-clés : philosophie morale ; intelligence artificielle ; éthique de l'intelligence artificielle ; roboéthique ; véhicules autonomes ; agents moraux artificiels ; moralité artificielle

The Implementation of Ethical Decision Procedures in Autonomous Systems: The Case of Autonomous Vehicles

Summary

The ethics of emerging forms of artificial intelligence has become a prolific subject in both academic and public spheres. A great deal of these concerns flow from the need to ensure that these technologies do not cause harm—physical, emotional or otherwise—to the human agents with which they will interact. In the literature, this challenge has been met with the creation of *artificial moral agents*: embodied or virtual forms of artificial intelligence whose decision procedures are constrained by explicit normative principles, requiring the implementation of what is commonly called *artificial morality* into these agents. To date, the types of reasoning structures and principles which inform artificial morality have been of two kinds: first, an ethically maximal vision of artificial morality which relies on the strict implementation of traditional moral theories such as Kantian deontology or Utilitarianism, and second, a more minimalist vision which applies stochastic AI techniques to large data sets of human moral preferences so as to illicit or intuit general principles and preferences for the design of artificial morality. Taken individually, each approach is unable to fully answer the challenge of producing inoffensive behavior in artificial moral agents, most especially since both forms are unable to strike a balance between the ideal set of constraints which morality imposes on one hand, and the types of constraints public acceptability imposes, on the other. We provide an alternative approach to the design of artificial morality, the *Ethical Valence Theory*, whose purpose is to accommodate this balance, and apply this approach to the case of autonomous vehicles.

Keywords : moral philosophy ; ethics ; artificial intelligence ; ethics of artificial intelligence ; roboethics ; machine ethics ; autonomous vehicles ; artificial moral agents ; artificial morality

SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE :
ED V – Concepts et langages
Maison de la Recherche, 28 rue Serpente, 75006 Paris, FRANCE

DISCIPLINE : Philosophie