UNIVERSITÀ DI SIENA 1240

Dipartimento di Economia Politica e Statistica

Dottorato in Economia

35° Ciclo

Coordinatore: Prof. Simone D'Alessandro

The Ethics of Artificial Intelligence from an Economics Perspective: Logical, theoretical, and legal discussions in autonomous vehicle dilemma

Settore scientifico disciplinare: SECS-P/01

Candidato
DAE-HYUN YOO
Università di Siena

Supervisore
Prof. NICOLA DIMITRI
Dipartimento di Economia Politica e Statistica

Anno accademico di conseguimento del titolo di Dottore di ricerca
2023

Abstract

The development of artificial intelligence (AI) systems and its impact on our societies raise ethical concerns. Ethical problems are not readily explainable to people. The persisting problem of autonomous vehicle (AV) ethics is how to handle ethical dilemma situations where the moral values of different people and ethical principles are in conflict. There is no consensus about what ethical choices should be made and which moral principles should be embedded to guide AV's decisions in ethical dilemma situations. The list of ethical principles and AI ethics guidelines say nothing about what to do when principles come into conflict with one another. There is few research about the tangible implementation of ethical values in the field of AI. Floating conclusion is used to conceptualize conflicting ethical principles that can be extended to AV's ethical dilemma. A floating conclusion approach in dilemma enables to implement conflicting moral values independently in AV settings. A new type of reasoner in the AV decision process is proposed to support the existence of two distinct types of AV. This reasoner reconciles the moral values and personal self-interest of the traffic participants by embedding different moral preferences to each of the two AV types. A static Bayesian game model is used to design incentives for a mechanism that addresses heterogeneous moral preferences in an AV decision dilemma, prevents human traffic participant's moral hazard behavior, and improves transportation efficiency in mixed traffic. No single but multiple ethical values allow policymakers to develop feasible, practical and effective mechanism designs for a smooth human-AI collaboration in the AI age. My dissertation explains the logic and ethical decision making in AI system meaningfully. It contributes to the body of ethics considerations in Human-AI interactions, specifically in the underexplored area where ethical principles and moral values of different participants in this interaction conflict with each other.

Acknowledgements

I am indebted to Prof. Nicola Dimitri for the many valuable comments and suggestions throughout my dissertation. I thank for his support and encouragement to complete. I thank Prof. dr. ir. Jan Broersen for the opportunity to visit Utrecht University, which first prompted me to work on logic to reason over conflicting values and principles. I thank Prof. Marija Slavkovik for the opportunity to conduct on Ethics of AI at University of Bergen. My thinking on the two different types of autonomous vehicle was influenced by conversations and thoughtful discussions about new two types of reasoners in an autonomous vehicle decision process. I also thank Dr. Michal Klincewicz for the many insightful discussions about liability in AV ethics at Tilburg University AI Special Interest Group (TAISIG). I especially thank Dr. Rune Nyrup for the valuable and thoughtful comments on the mechanism design of payoffs in game model at University of Cambridge. Most of the research was conducted while I was a visiting fellow in the Department of Information Science and Media Studies at University of Bergen and Leverhulme Centre for the Future of Intelligence (LCFI) at University of Cambridge. I also thank Prof. Vincenzo Valori, Prof. Ugo Pagano, and Prof. Sam Bowles for thoughtful comments at the annual meeting in Pontignano. I thank Prof. Ali Ozkes for the invitation to present at SKEMA Business School on Bayesian Game in the Human-Robot Society. I thank Prof. Prahlad Kasturi for the helpful advices. I am grateful for the many valuable comments and insightful discussions with participants of the Logic and AI at University of Bergen, Dutch Logic PhD Day 2022 at Utrecht University and of the workshop Ethics of Public Robots and AI at SKEMA Business School in Paris. I am grateful to Petra for emotional support and encouragement to finish my writing. I am also grateful to my colleagues in 35th cycle for the critical discussions about many issues during the coursework. Lastly, I thank my parents and sister. They are always on my side and their belief in me makes me go through all the difficult situations. I am dedicating my doctoral dissertation to my beloved late grandparents who are always in my memory.

Table of Contents

Tables

Figures

Graph

Abbreviation

AI	Artificial intelligence
AMA	Artificial moral agent
AMAs	Artificial moral agents
AV	Autonomous vehicle
AVs	Autonomous vehicles
AVI	Type I Autonomous Vehicle
AVII	Type II Autonomous Vehicle
DAV	Default Autonomous Vehicle
EC	European Commission
ECB	The European Central Bank
EPIC	Electronic Privacy Information Center
EU	European Union
FED	The Federal Reserve System
IEEE	Institute for Electrical and Electronics Engineers
NHTSA	The National Highway Traffic Safety Administration
OECD	Organisation for Economic Cooperation and Development
SAE	The Society of Automotive Engineers
UK	United Kingdom
UN	United Nations

Introduction

The development of artificial intelligence (AI henceforth) systems and its impact on our societies raise concerns about the ethical issues of AI. People usually regard it as good and bad in general. But there is something unique about AI ethics in the way AI and automated decision-making systems have led to some intended or unintended harmful outcomes.

The general ethical principles in AI are beneficence, autonomy, justice, and explicability. Beneficence means do good and no harm which can be regarded as non-maleficence. Autonomy preserves human agency and justice means to be fair. These are similar to bioethics. Explicability is the specific ethical problem raised by AI due to its technology[1]. For AI to be ethical, trustworthy, and transparent, it must be explainable, especially when making ethically critical decisions, such as life and death. Explicability is a series of processes enabling humans to better understand how algorithms are used in decision making and what factors lead to AI's decisions and predictions. It is a necessary condition for responsible and accountable behavior and decisions of AI. AI systems' decisions have major impacts on people, so we must be able to understand the underlying reasons for them. That is, AI and its decision-making need to be explainable. AI's explicability is referred as a basic ethical criterion, among others, for the acceptability of AI decision-making. However, many decisions made by autonomous AI systems are not readily explainable to people. This is also called the problem of opacity (Gordon & Nyholm, 2021). Grinbaum et al. (2017) emphasize that researchers must deal with the problem of opacity of the robot's decisions by focusing on the increased capacities for autonomous decision making and on the associated risks, such as the difficulties in programming moral judgment.

When people are concerned about ethics of AI, they often question about its' autonomous ethical decision making in dilemmas. The life-or-death relevant judgment involves ethical considerations such as driverless car case. As the information available to automated systems gets richer, the ethical dilemmas it confronts will also grow more complex. With the increasing complexity of such systems, the evaluation of conflicts between values

[1] Machine learning, one of bottom-up and traditional engineering approaches of testing and refining intelligent systems (Wallach & Allen, 2009). It is a main source of unexplainable AI, called the Black Box problem. This may be one of core characteristics of AI in a way that the logic or intent behind the output of systems can often be extremely hard to explain: the adaptiveness of the technology (UK, 2022). Therefore, the United Kingdom government propose that "the government should not set out a universally applicable definition of AI. Instead, we will set out the core characteristics and capabilities of AI and guide regulators to set out more detailed definitions at the level of application." (UK 2022). The United Kingdom government's pro-innovation approach is discussed in Section 6, Chapter 3, The role of government for ethical AI.

becomes increasingly problematic. In the face of such uncertainty, there is a need for autonomous systems to weigh "risks" against "values". Wallach and Allen (2009) mention that it needs to understand which values AI can and should promote as for the built-in ethical constraints or for the capable of following the ethical standards. Yet, there is no consensus about what ethical choices should be made and which moral principles should be embedded in the AI system (Lin et al., 2012; Lin et al., 2017; Wallach & Allen, 2009). Scholars are debating which moral doctrines and principles should be embedded to control and guide AI's decisions in ethical situations.

People normally take moral decision making for granted in our daily lives. In the case of ethical questions, different people have different opinions about what's ethical and what's not. For example, in the ethical dilemma cases of autonomous vehicle (AV henceforth), views on whose safety should be prioritized differ. Many ethicists have agreed that an action should only be considered moral if it stems from a certain state of the agent's mind, a certain quality of intention, purpose, motive, or disposition. Some argue whether AI should be considered as moral agents or not. It has been argued that machines should never be allowed to make life-or-death decisions even when those machines are used in war (Scharre, 2017). Etzioni and Etzioni (2017) point out that an autonomous machine, including driverless cars, cannot make ethical decisions and further claim that the relevant ethical challenges can be addressed by the ethical choices made by a human. Grinbaum et al. (2017) state that an autonomous robot cannot be equipped with completely adequate ethical rules. van de Poel (2020) claims that we need to treat an AI system as a value-laden sociotechnical system without conceiving it as a moral agent. In the case of collision-avoidance algorithms with pedestrians in AV, van de Poel (2020) suggests that upon embedding values in autonomous systems, we need to consider a socially acceptable way to pass a pedestrian; socially acceptable to the community in which the system will be deployed.

The most interesting question is what if some of ethical principles conflict with each other when it comes to AI ethics, how should we choose and implement it in AI systems? What could be a general standard for measuring moral value and for making hard moral decisions? Especially in the specific ethical dilemma case, which ethical standard and method would be better to adopt for the AI ethics? Do we want to embed practical human ethical values or ideal ethical values in AI systems? Finding the right set of ethical considerations as constraints and the right formula for resolving conflicts is another challenge for making an ethical AI.

People want to have a human-liked or human-centred AI[2] that should not get stuck when it encounters an ethical dilemma. The human-centred AI would learn how humans make decisions about ethical problems and develop moral reasoning. If then, what about human's ethical decision making in ethical dilemma cases? What if humans were in an ethical dilemma situation, what would they do in reality? How would humans, in practice, make ethical decisions in specific ethical dilemma cases? For example, what should people and AI choose between driving into a child or into a wall to save the child's life but potentially killing its passenger? What kind of ethical constraints should driverless cars have built-in in their AI system? If so, how should they be determined? Such hard questions as in the ethical dilemma cases are important for AI ethics. Because it would be one way to test ethical AI systems by exercising certain degrees of skepticism and asking tough questions. Floridi et al. (2018) encourage to explore "what-if" scenarios by testing the AI system prior to fully development. Many of the real-world moral codes do not prioritize all the rules and are therefore subject to conflicts between them (Wallach & Allen, 2009, p. 93). The literature at present also emphasize the importance of solving the multiple ethical preferences dilemma (Bonnefon, 2016; Bonnefon et al., 2016; Bonnefon et al., 2019; Whittlestone et al., 2019), while recognizing the challenge of designing autonomous machines with our moral values embedded (Coeckelbergh, 2019; Lin, 2013; Lin et al., 2017; Wallach & Allen, 2009; Wallach et al., 2010).

If humans want to program the values inherent in the humans' ethical behavior, it is better to understand human's ethical behavior. Yet, there is few research about how to resolve conflict between two or more opposing ethical principles and address it as the ethical decision-making problem for humans. In addition, to be able to have a moral decision making of AI, we need to recognize what kinds of ethical decisions can and cannot be made by its systems. AI's ethical reasoning does not need to be broad as something like "friendly" or "trustfulness". At least, it should manage ethical uncertainty and ambiguity by involving the specific tasks with being friendly and trustful in the given role. These demand human moral decision making, particularly in ethical dilemma cases that have not yet been analyzed.

The ethics of AI could be considered as the way to resolve conflicts between two or more conflicting principles. The AI ethics could be also addressed as the ethical decision-

[2] AI systems should be designed in a way that respects the rule of law, human rights, democratic values, throughout the AI system lifecycle, and they should include appropriate safeguards – for example, enabling human intervention when necessary – to ensure a fair and just society (EPIC, 2023; OECD, 2023).
Human-centred values is a principle for responsible and trustworthy AI such that "AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art." (OECD, 2023).

making problem for humans in specified ethical dilemmas. Wallach and Allen (2009) claim that the project of building artificial moral agents (AMAs) highlights the need for a richer understanding of human morality. That is, building AMAs forces us to take a particularly comprehensive approach to ethical decision making of humans. Meeting these challenges requires understanding of human ethical behavior. Programmers and developers of AI may embed the ethical principles by separating deontological ethics (rules, laws, and other statements of ethical duty) as the default and utilitarianism ethics as adjustable local ethics that explains other left areas for different humans' ethical preferences across the country in the AI system. Yet, the list of ethical principles and ethical guidelines are quite abstract to implement (EPIC, 2023; OECD, 2019, 2023; *The Universal Guidelines for Artificial Intelligence*). For every moral principle, there appear to be moral trade-offs. The key ethical values and principles, for example beneficence, autonomy, justice, and explicability may conflict with each other to embed in AI systems. Then, how should we embed it in AI systems? What if more than two ethical values and principles conflict with each other[3], what is human's ethical and moral choice? Are we good at making moral decisions? We have many questions but few answers. There are a few concrete proposals. The lack of clarity on human ethical choice problems increases the prospect of political influence in determining whether new technologies are regulated. Fear of possible bad consequences is just one concern on the minds of politicians. We must set reasonable goals and work reasonably to attain them not in principle, but in practice. As Haidt (2012) says, we are very good at seeing through our opponents' moral rationalizations[4], but we need to get better at seeing our own. Wallach and Allen (2009) point out the importance of understanding humans' ethical decision-making to build an artificial moral agent (AMA):

> The process of designing (ro)bots capable of distinguishing right from wrong reveals as much about human ethical decision making as about AI (p. 215).

The ethical issues of AI teach us something about humans and human societies by revealing our ethical choices and preferences in ethical dilemmas. Understanding humans'

[3] It refers to situations in which a choice of one principle leads to infringing some other available principles that cannot be avoided. That is, available choices (embedded with principles or values) come into conflict with each other. So, when an individual decides to choose one, it must sacrifice another choice.
The definitions of dilemma and ethical dilemma in the autonomous vehicle ethics are discussed in Section 3.1.

[4] I do not take a position of AI as human's "opponent" for us. However, when people discuss about the ethics of AI, they often not only criticize bad design and programmer's failures to build in adequate safeguards, but also request very strict ethical standards to AI without addressing our own ethical choices in dilemma cases.

ethical decision making in a situation where available ethical values and choices come into conflict with one another is a requisite to embed ethical values and principles into AI systems. Humans program algorithms in AI systems. The big data in AI represents humans' behaviors. The human-centered AI would learn how humans make decisions about ethical problems and develop moral reasoning.

Following these observations, I am motivated by the problem of handling ethical dilemma situations where the moral values of different people and ethical principles are in conflict. The remainder of my dissertation is structured as follows.

First in Chapter 1, floating conclusion is applied as logical tool and skeptical reason over conflicting economic principles and goals as *ex ante*. The floating conclusion approach justifies the conclusion from extensions associated with conflicting propositions. In Section 1, the conflicting principles and goals in economics are described. In Section 2, a floating conclusion approach is introduced, and its structure is analyzed. The validation of this logical approach in economics can be extended to ethical dilemma cases where ethical principles and moral values conflict with each other in the AI system.

In Chapter 2, game models are adopted in an autonomous vehicle decision dilemma as the specific case of AI ethics. The persisting problem of AV ethics is how to embed heterogeneous moral preferences in the context of protecting different traffic participants to guide the decisions of an AV in dilemma situations. The difficulty of designing moral and ethical AV is discussed in Section 3. In Section 4, a new type of reasoner in the AV decision process to identify conflicts in driverless cars where the different moral preferences come into conflict with each other is taken. It supports the existence of two different types of AV and reconciles personal self-interest of traffic participants by embedding different moral preferences to each of the two AV types. Then, static Nash and Bayesian games to analyze strategic interactions between pedestrian and two different types of AV are applied in Section 5. A game-theoretic approach is an alternative to address ethical issues by designing incentives for the AVs and traffic participants that make ethically aligned behavior and rational choice in a dilemma. This game theoretic approach in AV dilemma also shows to prevent human traffic participant's moral hazard behavior and improve transportation efficiency in mixed traffic.

Lastly in Chapter 3, the role of government for the ethical AI system is discussed. In Section 6, the current AI ethics guidelines and policies are discussed. The applications of floating conclusion approach as a logical approach and Bayesian game model in AV dilemma are in Section 7. The policy implication is derived to enable policymakers to develop feasible, practical and effective mechanism designs for a smooth human-AI collaboration in the AI age.

Finally, in Section 8, liability and responsibility issues such as who should be responsible for the AV's unintended and intended choice that causes harm to humans, and how to distribute responsibility are discussed as further research topics. A dynamic consistency of static Bayesian Nash equilibrium is also discussed as further research and I conclude.

My doctoral dissertation answers how to make ethics computable for aiming to steer human-like AI by bridging the gap between the very abstract ethical values and principles to concrete technical implementations. It contributes to the explicability of AI in ethical dilemmas, because the logic and ethical decision making in AI systems can be meaningfully explained in an intelligible way. It also contributes to the body of ethics considerations in Human-AI interactions, specifically in the underexplored area where ethical principles and moral values of different participants conflict with each other. This supports the human-centric society for a smooth human-AI collaboration in the AI era. My doctoral dissertation also improves our application of ethical principles and moral values in the AI age and contributes to developing a framework to enhance the explicability of AI systems.

I. The Logical Approach in Dilemma
The Skeptical Reason over Conflicting Economic Principles and Goals

1. The Conflicting Principles and Goals in Economics

The economic policy tools are designed and implemented to resolve a different level and dimension of conflicts in society. The traditional economic approach in addressing conflicting economic principles, such as efficiency versus equity and evaluating them as *ex post* is based on the individual indirect utility function, its welfare, and a social welfare function. The pursuit of efficiency goals impairs equity and vice versa. Policymaker is struggling to handle conflicting economic principles in dilemmas and pursues policy tools.

1.1 Introduction

The recent rise of overall living cost, due to the supply and energy shocks from the ongoing Russian and Ukraine conflict leads central banks to raise interest rates in developed countries. There are growing concerns about their monetary policy that could have some serious economic downturn. The World Bank and the UN have warned that it could trigger a global recession, especially the developing country's bankruptcy as chain reactions such as in Sri Lanka, Ecuador, Peru, and Tunisia. Nevertheless, the key developed countries' central banks have responded to rising inflation with aggressive interest rate hikes even in the risk of recession in domestic. The central bank of the United States (Fed) hikes the federal funds rate as the benchmark interest rate again by 75 basis points to a range of 3.75% to 4%, its highest level since 2008. According to the Federal Open Market Committee's statement, "ongoing increases" of interest rates will likely be needed to a level that is "sufficiently restrictive to return inflation to 2% over time." This indicates how policymakers' decision firmly committed to curb inflation as stated (Fed, 2022). The Bank of England raises UK interest rates to 3% in more than three decades. "Output has already started to contract and will go on falling for the next two years – the longest recession of modern times. Unemployment will almost double, with the jobless rate rising from 3.5% to almost 6.5%. Inflation will fall from 10.1% to well below its 2% target over the next two years." (Guardian, 2022). The European Central Bank (ECB) decided to raise the three key ECB interest rates by 75 basis points, third rate increase in a row. They expects to raise interest rates further, to ensure the timely return of inflation to its 2% medium-term inflation target ((ECB), 2022). However, such monetary policy of raising interest rates to lower inflation causes higher unemployment rate, at least in the short run. The relationship between inflation and unemployment is known as the *Philips Curve*. Two main arguments about this relationship developed in the mid-twentieth century. The first viewed the

Philips curve as a dilemma that forces policymakers to choose between price stability and full employment. The second does not see a long-run conflict between these two goals (Schoder, 2018).

1.2 Definition of conflict

Conflict is a condition in which there is a perceived difference in interests or goals. It can be a type of competitive behavior between individuals or groups that they fight over it. The dictionary definitions define the conflict as a *disagreement* or *dispute* between people. When two or more people or groups think they have conflicting interests or goals, tensions arise. It is a natural consequence of people's different ideas, beliefs, attitudes, and perceptions.

Conflicted interests represent the dilemmas in which a choice of one choice (embedded with principle or value) leads to infringing some other available choices that cannot be avoided. That is, available principles come into conflict with each other, so when an individual decides to choose one, it must sacrifice another principle.

Much of the economic theory is concerned with the problem of rationalizing the combined and interdependent effects of a set of independent individual choices based upon conflicting preferences (Bower, 1965). The theories of oligopoly and monopoly are concerned with the effects of conflict when the number of opposing firms is small. Game theory is an economic analysis tool for conflict resolution. It addresses conflict and tension between agents and explains how to lessen or resolve conflict between them. On the other hand, conflict challenges policymakers as the third party to resolve a dilemma situation.

1.3 Conflicting principles and goals

The effective policy design is complicated because conflicting plural principles and the socially desirable tradeoff among them must be taken into consideration to ensure an optimal and sustainable outcome. Such conflicting principles and goals in economics are shown as follows.

1.3.1 Supply and Demand

The basic demand and supply economic model represent two conflicting principles in the context of market participants as buyers and sellers. For example, buyers always want to pay less but sellers want to get paid more. Quantity demanded is the amount of goods that buyers are willing and able to buy at given prices. *Law of Demand* says that the quantity demanded of goods falls (rises) when the price of the goods rises (falls), other things equal. On the other hand, quantity supplied is the amount of goods that sellers are willing and able to sell

at given prices. *Law of supply* says that the quantity supplied of goods rises (falls) when the price of the goods rises (falls), other things equal. These conflicting principles are graphed as the *downward* sloping market demand curve and *upward* sloping market supply curve.

1.3.2 Efficiency and Equity

There are two conflicting goals, such as efficiency versus equity in economics. The equity versus efficiency dilemma is the main issue in the political debate. The evidence of widening economic inequalities raises equity to be a priority concern among policymakers. Although the general aim of reducing economic inequalities appears uncontroversial, there is no consensus on how to deal with policies that may cause a conflict between the goals of equity and efficiency. It challenges policymakers to deal with policies that may cause a conflict between the goals of equity and efficiency. In taxation, these two conflicting economic and ethical principles such as equalitarianism versus utilitarianism are observed. Equalitarian ethics is used to defend progressive taxation such that a flat percentage tax would be a disproportionate burden for people with low incomes. The U.S. income tax system is an example of progressive tax scheme that pursue equity. It imposes a higher tax rate on high-income earners than on those with a lower income. On the contrary, utilitarian ethics provides a defense for equalitarian measures by using the principle of maximum utility. A sales tax imposes the same percentage rate on products or goods purchased, regardless of the buyer's income. It is considered as regressive tax scheme to pursue efficiency but impair equity. Because it takes a larger portion of disposable income from low-income earners than from high-income earners.

2. Floating conclusion as Skeptical reason

A floating conclusion is a logical tool to reason over conflicting propositions as *ex ante*. It explicitly describes dilemma situations and provides a framework to conceptualize conflicting ethical principles.

2.1 Definition and Structure

Skeptical reason is a logical tool to reason conflicting principles or values (Bonnefon, 2004; Horty, 2002). The floating conclusions approach compares two conflicting propositions and justifies the conclusion from those conflicting propositions. Floating conclusions are supported in each extension associated with a knowledge base, but only by different arguments (Horty, 2002, p. 61). The author describes it as "a situation in which two sources of information,

or reasons, support a common conclusion. But it also undermines each other, and therefore undermine the support that each provides for the common conclusion." (Horty, 2002, p. 68). According to Bonnefon (2004), *D* is called a floating conclusion within the below argument frame in Table 1:

A and B If A then C and D, If B then not C and D,

Table 1 The definition of floating conclusion

Horty (2002) and Bonnefon (2004)'s definitions of floating conclusions is structured as the simple form[5] in Figure 1 and as the extensive form in Figure 1.1.

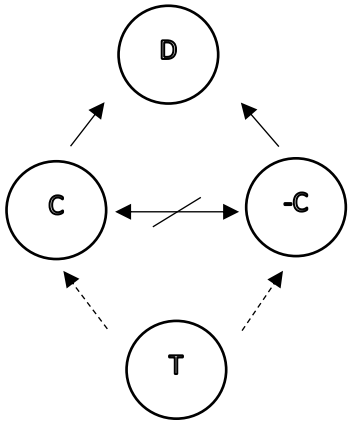


Figure 1 The simple structure

T is the observation, in which two different sources of information or reasons *A* and *B* are included. Two conflicting propositions are presented by C and $-C$, where $-C$ represents *Not C*. These conflicting propositions can be derived from *A* and *B*, respectively. *D* is a floating conclusion from conflicting propositions, and it can be accepted or not accepted. An *acceptable* floating conclusion does not mean to justify conflicts, but a conclusion from conflicting propositions. An *unacceptable* floating conclusion fails to justify a conclusion from conflicting propositions. Arrows indicate the logical flow such that broken arrows are default implications and solid ones are ordinary logical implication.[6]

[5] Source: On Floating Conclusions (Schuster & Broersen, 2022)

[6] Horty (2002) and Schuster and Broersen (2022) distinguish between ordinary logical implications and "default" implications. However, all conditionals are default assertions (Bonnefon, 2004).

The extensive structure of floating conclusions captures the higher dimension of conflicts in Figure 2.1. For example, two agencies – the government (A) and the central bank (B) exist whose principles conflict with each other; the government's *expansionary* fiscal policy (C) versus the central bank's *contractionary* monetary policy ($-C$). Yet, promoting sustainable economic growth as a common goal (D) from two conflicting policies would be justified and accepted.

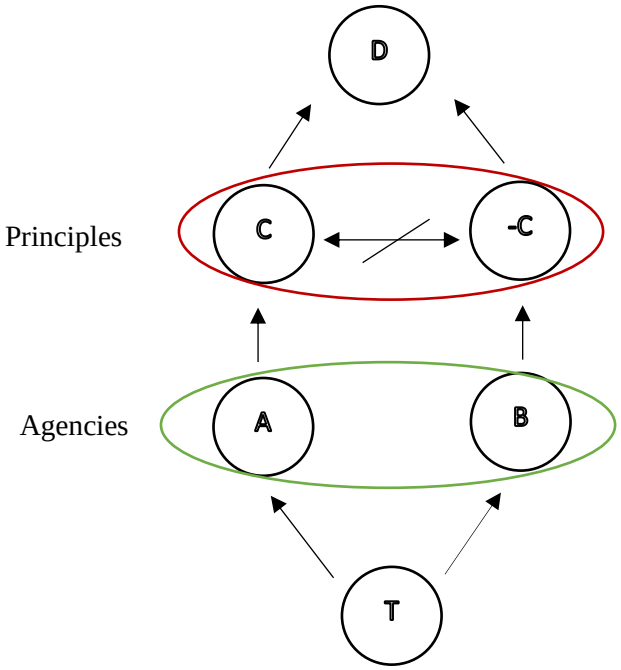


Figure 1.1 The extensive structure

The arrow ($\rightarrow$) can be interpreted as the subset ($\subseteq$) in the context of set theory. $A \rightarrow C$ implies A is a subset of C and set A is included in set C: $A \subseteq C$. $C \rightarrow D$ implies C is a subset of D and set C is included in set D: $C \subseteq D$. Therefore, $A \rightarrow C \rightarrow D$ means that $A \subseteq C \subseteq D$. Likewise, $B \rightarrow -C \rightarrow D$ implies $B \subseteq -C \subseteq D$. D is a floating conclusion, *if and only if* $D = \Omega$ which would be impossible. Because D cannot be Ω, where Ω is defined to be the set of all finite ordinals. Instead, the interpretation of the arrows of logical and default implications in floating conclusion makes D as a floating conclusion possible as follows.

$$A \rightarrow (C \cap D)$$
$$B \rightarrow (-C \cap D)$$

Then, D is a floating conclusion from conflicting propositions C and $-C$, *regardless of* $D = \Omega$.

2.2 Floating conclusions approach in economics

The evidence of persisting and widening economic inequalities raises equity to be a priority concern among policymakers. Although the general aim of reducing economic

inequalities appears uncontroversial, there is no consensus on how to deal with policies that may cause a conflict between the goals of equity and efficiency. That is, those may increase economic inequalities or improve fairness while decreasing efficiency. The equity versus efficiency dilemma is the main issue in the political debate.

2.2.1 Acceptable and Unacceptable Floating Conclusions

An acceptable floating conclusions is justified as a common conclusion from extensions associated with conflicting propositions. It indicates that multiple conflicting principles or goals can be implemented with an acceptable common conclusion from those conflicting propositions.

2.2.1.1 The long run goal

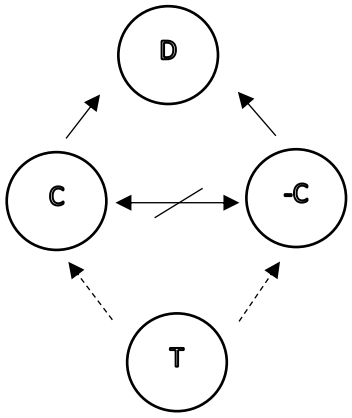


Figure 2 An acceptable floating conclusion
The central bank's monetary policy in the long run

The central bank conducts monetary policy to lower inflation. However, it undermines promoting maximum employment and causes high unemployment. That is, the central bank's two short-run goals – promoting maximum employment and remaining at low inflation, conflict with each other. This can be structured in the simple form of floating conclusions approach in Figure 2, where

$$\begin{cases} T: \text{The monetary policy tool} \\ c: \text{Price stability in the short run} \\ -c: \text{Maximum employment in the short run} \\ D: \text{The long run goal} \end{cases}$$

T is the current ongoing central bank's monetary policy of raising interest rates. Raising interest rates *implies* not only to curb inflation (C) but also to hinder employment $(-C)$, at least in the

short run. Yet, the steady economic growth in the long-run is justified as an acceptable floating conclusion (*D*) from these two conflicting short-run goals.

2.2.1.2 The equilibrium

In a free and unregulated market system, market forces establish equilibrium prices and quantities of goods and services. Equilibrium refers to a situation where quantity supplied equals quantity demanded. At market equilibrium, buyers and sellers receive benefits from taking part in the market. These benefits are considered as consumer surplus and producer surplus. However, as the price of goods falls, consumer surplus increases but producer surplus falls.

While equilibrium conditions may be efficient, it may also be true that not everyone is satisfied. For example, when policymakers believe that the market price is unfair to buyers or sellers, they control prices. The government intervenes in the markets to protect the buyers or sellers. In this case, the equilibrium price or quantity of goods and services is *ineffective* and *not accepted* in the market. Therefore, it could be an *unacceptable* floating conclusion.[7]

2.3 Results

Floating conclusion is a logical approach to reason conflicting principles or values. It is a knowledge representation and reasoning tool to formulate conflicting propositions. The structure of floating conclusion addresses dimension of conflicts in principles or goals. An acceptable floating conclusion is justified as a common conclusion from extensions associated with conflicting propositions. A floating conclusion is acceptable, if it is *the long run* goal from extensions associated with two conflicting *short run* goals. The equilibrium price and quantity are acceptable floating conclusions if it is a *desirable* and *final* goal. This approach provides useful guidance for policymakers to design effective policy tools with an acceptable balance between conflicting policy goals. It implies that policymakers consider conflicting economic principles or policy goals independently and accept them as extensions to support a common conclusion from those conflicting ones.

[7] For example, price ceiling is a legal maximum on the price of a good or service at which a good can be sold. It can be set above or below the market equilibrium price. For a price ceiling to be effective, it must be set below the equilibrium price to protect the buyers and it is called a *binding* price ceiling. In this case, the equilibrium price is *ineffective*, therefore, it is an unaccepted market price.

2.4 Discussion

People discuss how to embed human ethical principles and moral values to guide autonomous agent's decisions. People explain how it is challenging the design of autonomous machines with our moral values in machine ethics. The ethical issues of AI teach us something about humans and human societies by revealing our ethical choices and preferences in ethical dilemmas. Before we integrate our moral values into AI systems, a better understanding of how people make ethical decisions and people's ethical decision-making process is a precedent of discussion about the ethics of AI.

Ethics is incomplete by numbers alone; rights, duties, conflicting values, and other factors often come into play. Different ethical theories such as deontological, utilitarian, virtue ethics differ on the importance of moral justification and provide different features of actions and justify outcomes in ethically challenging situations. Humans' ethical reasoning can be thought as how different schools of philosophers think about what a single general principle should be – utilitarianism and deontology in top-down approach[8] for ethical AI systems.
Gawronski and Beer (2017) explains the difference in responses to moral dilemmas between utilitarianism and deontology. According to the principle of utilitarianism, the moral status of an action depends on its outcomes, more specifically its consequences for overall well-being (outcome-based morality). If a particular action increases overall well-being in a given situation, it is deemed morally acceptable from a utilitarian view. Yet, if the same action decreases overall well-being in a different situation, it is deemed morally unacceptable in that situation. Utilitarianism allows individual agents to have moral preferences in utility functions and to act in their interests. The utilitarian's approach to make ethical decision is about maximizing the aggregate of utility, which is a measure of happiness and welfare.

[8] Top-down approach to moral judgment defines what is moral and not moral. This approach takes a specified ethical theory such as consequentialism, deontology and lists of virtues, and it plays an important role not only for evaluating the morality of actions, but helping people sort out cases where moral intuitions are unclear. Its sources are variety by religion, culture, and philosophy. The Golden Rule, the Ten Commandments, legal and professional codes, Asimov's Three Laws of Robotics, and Kant's categorial imperative are examples of this approach. Professional ethicists know that ethical theories, in particular top-down approaches cannot provide real-time decision procedures. Instead, many of them see the ethical issues as aimed at justifying ethical decisions within a comprehensive framework. (Wallach & Allen, 2009).
In a context of driverless cars, ethical principles are programmed into the car's guidance system with a general moral philosophy. So, the car will be able to make ethical choices based on the moral philosophy that was implanted onto its AI program. (Millar, 2014a).
The utilitarian and deontological approaches to ethics have been the most widely considered as to the development of ethical AI systems within a top-down approach. The distinction between utilitarianism and deontology is a prevailing framework to conceptualize moral judgment. Therefore, these two moral principles will be considered as conflicting principles in the context of AI ethics thereafter in my dissertation.

Jeremy Bentham developed utilitarian views regarding morality as an objective by getting away from dependence on hard-to-justify lists of duties or individual intuitions about what is right or wrong. Instead, Bentham developed a method of evaluating situations quantitatively. Bentham assigned numbers to the goods and harms that resulted from actions and quantitative measured utility that allows for a simple decision rule: choosing the action that results in the highest total utility than not choosing it. John Stuart Mill, the most famous utilitarian of the nineteenth century, argued that actions are morally independent of the agent's motivations.

According to Allen Newell, people do practice with bounded morality[9] given abstracts and broad ethical principles and rules. Ethical heuristics is a concept of following rules that are expected to increase local utility rather than analyze all the consequences of one's possible actions. So people may choose between them on the basis of maximizing their local utility. These concepts could abbreviate the relevant ethical choices of the domain in the AI system and evaluate certain choice and its outcomes as more valuable than others. It could also enable us to explain why some ethical considerations could be considered and adopted in AI system. They play an important role in translating broad and abstract ethical values and principles into the weights on particular moral considerations. However, moral philosophers think that the footbridge dilemma[10] highlights a fundamental flaw in utilitarian thinking and do not follow this approach. The philosophical challenge to utilitarianism is whether moral and ethical values could be translated into a single metric and its aggregate.

Greene (2014) claims that the values behind utilitarianism are our true moral common ground by defining happiness as the follow:

> Happiness is what matters, and everyone's happiness counts the same. This does not mean that everyone gets to be equally happy, but it does mean that no one's happiness is inherently more valuable than anyone else (p. 170).

It implies that the utilitarian approach to ethics counts everyone's happiness equally but achieving its success could occur unequally. If this inequality could not be justified, things would be worse overall. Because we know that the ethical principle or the standard of happiness

[9] It is the concept of "bounded rationality" in ethics. It is a limited set of ethical values or principles in one's moral and ethical decision making.

[10] The Footbridge, a large man can be pushed in front of the trolley. The large man will die, but his body will stop the trolley before it can kill the five workers on the track (Awad et al. 2020).

is different things for different people. Even if someone claims that the ethical principle is the same thing for everyone, different things make different people happy and unhappy.

In contrast to the utilitarian judgments, the deontological approach to ethics regards morality as a duty or a moral rule that ought to be followed. It is about following universal norms that prescribe what people ought to do, how they should behave, and what is right or wrong. Deontologists, particularly Kantian ethicists claim that there is more to rights and wrong than maximizing overall happiness. The principle of deontology emphasizes the situation-independent status of moral norms (rule-based morality). According to the principle of deontology, a given action is morally acceptable if it is consistent with relevant moral norms, but it is morally unacceptable if it is inconsistent with relevant moral norms. Van Staveren (2007) explains how deontology is applied in economics. Deontological notions such as 'rights', 'equality', and 'norms' appear more prominently in heterodox traditions such as political economy, institutional economics, and socioeconomics. The common ground is a recognition that rights, moral rules, and norms affect economic behaviour as constraints on choice. The examples of deontological ethics in economics are anti-trust laws, prohibitions on insider trading, and anti-dumping regulations. That is, there are universal moral rules that are enforced through legal measures.

The challenging part is whether the relevant actions are allowed by the rules, especially in the ethical dilemma. Deontological ethics has no criterion for dealing with conflicting rules (Van Staveren, 2007). In addition, not all moral problems can be solved by rules. Because human life is too complex to be reduced to a set of rights and duties. For example, the rules of the road constrain the behavior of human drivers to minimize the risk of injury and death and to promote traffic flow. However, the rules alone may be insufficient to define the correct behavior for the agent in all circumstances. Prakken (2017) points that some actions are technically illegal, but acceptable by social norms (such as driving slightly above the speed limit to match surrounding vehicles), or vice-versa (driving below the speed limit to an extent which inconveniences and angers human drivers). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems discusses the importance of social and moral norms to embed values into autonomous systems. These norms depend on the community in which the system will be deployed, and they might change over time. In the example of collision-avoidance algorithms with pedestrians, what is considered as a socially acceptable way to pass a pedestrian will depend on where the AI system is operating in which country and which city (IEEE, 2018). Wallach and Allen (2009) discuss the case of AV having to break the traffic laws to avoid an accident. Prioritizing rules may minimize conflicts between the available actions. Yet, many of the real world moral codes

do not prioritize all the ethical rules and are subject to conflicts between them (Wallach & Allen, 2009, p. 93). Wallach and Allen (2009) state that even single rules can lead to deadlock in a case when two humans give contradictory commands or face conflicting values. Conflicts can cause a deadlock that we do not want to have in AI systems.

AI with the top-down capacity to evaluate ethical choices and its immediate outcomes would be capable of selecting the actions that both meet its goals and fall within acceptable social norms. It could be more effective to embed ethical principles than bottom-up approach[11] in intelligent systems in general. The programmer may be able to anticipate the possible courses of action and provide rules that lead to the desired outcome. On the other hand, the programmer may build a more open-ended system that gathers information, attempts to predict the consequences of its actions, and customizes a response to the challenge with a bottom-up approach. However, both approach face challenges not only the prospect for implementing ethical rules as formal decision algorithms, but also what to do when ethical principles conflict. There is always the prospect that a learning system will acquire knowledge that conflicts directly with its in-built restraints. When there are conflicts between virtues[12] or incomplete lists of virtues, it would be difficult to program virtues into AI even with top-down and bottom-up approaches. Given the difficulties in establishing suitable and widely-accepted ethical codes to form the basis for ethical AI systems, it requires an ethical AI to manage dilemma situations in which moral principles conflict. However, we currently have no consensus about which moral doctrines and principles should be embedded to control and guide autonomous agent's decisions in ethical dilemma situations. We need to find a way to steer between many moral considerations that impinge on each other and come up with one that balances those considerations as well as possible. Understanding humans' reason over conflicting ethical

[11] This approach is out of coverage in my dissertation. However, briefly speaking, it is traditional engineering approaches of testing and refining intelligent systems. In a context of ethical decisions, machines are expected to learn how to render ethical decisions through observations of human behavior in actual situations without being taught any formal rules or being equipped with any moral philosophy. For example, driverless cars could learn from the ethical decisions of millions of human drivers as a sort of drawing on the wisdom of the crowds. But this may lead cars to acquire some rather unethical preferences too. (Millar, 2014a).

[12] Virtue ethics, the 3rd element of moral theory, emerges as the difficulties of applying general moral theories in a top-down approach. It says that the discussions of ethics are not just about rights (deontology) and welfare (utilitarian). According to virtue theory, morally good actions flow from the cultivation of good character, which consist in the realization of specific virtues. Virtue theorists emphasizes the importance of developing character or good habits rather than focusing on rules or consequences. While virtue ethics provides rich intellectual resources for philosophical reflections about character, these reflections are likely to remain relatively remote from the daily ethical reason practices. Furthermore, it is unlikely that the virtues can be clearly divided into top-down and bottom-up approaches. (Wallach and Allen, 2009).

principles is a requisite to align moral algorithms with human values and embed appropriate moral values into AI system.

Lin et al. (2017) claim that

> A combination of philosophical moral theories and analogical reasoning is the most promise way to engage in moral reasoning for engineering software (pp. 244-259).

Floating conclusion is a reasoning tool to formulate conflicting propositions. This skeptical reasoning is pragmatic approach to help to understand how to handle conflicting principles or values in dilemma situations. It is well suited to the individual choice setting for decision making over conflicting ones and explicitly describes dilemma situations.

A floating conclusion approach provides a framework to conceptualize conflicting ethical principles that can be extended to autonomous vehicle's moral dilemma cases where ethical principles and moral values conflict with each other. In particular, an acceptable floating conclusion from extensions associated with conflicting principles or values enables to embed conflicting ones independently and accept them as extensions to support a common conclusion in AV settings. For example, manufacturers embed the conflicting ethical principles separately with an acceptable common conclusion and integrate the interests of different parties involved in AVs environment. That is, they can produce different types of ethical driverless cars with both utilitarian and deontologist principle in their AVs. They have an incentive to advertise their operating ethical principles and it enhances transparency. This contributes to the explicability of AI in ethical dilemmas. To the best of my knowledge, no research has been conducted to address ethical dilemmas of AV with reasoning. The detailed analysis of floating conclusions approach in AV dilemma is discussed in Section 7.1.

II. The Game Theoretic Approach in Autonomous Vehicle Dilemma
A Static Bayesian Game in the Human-AI Society:
The interaction between pedestrian and two types of autonomous vehicle

Artificial intelligence (AI henceforth) systems are deployed in various fields of use in which they offer context-specific benefits and challenges. For instance, in transportation settings, autonomous vehicles[13] (AVs henceforth) have benefits to improve traffic safety and efficiency, e.g., use of fuel (Bonnefon, 2016; OECD, 2019). However, concerns are raised about the possible accidents that could occur and the ethics of crashes with self-driving cars that society urgently needs to deal with. For example, a pedestrian was hit and killed by an experimental self-driving car in March 2018 (BBC, 2020). The AI system in the self-driving car involved in the accident has not been successful in adequately classifying the object that suddenly appeared in its path.

When people are concerned about the ethics of AV, they often question how AVs should be programmed to handle the unpleasant life-and-death tradeoff between pedestrians and passengers or choose between two harms in cases when inflicting some harm cannot be avoided (Bonnefon, 2016; Bonnefon et al., 2016; Bonnefon et al., 2019; Coeckelbergh, 2019; Dignum, 2018; Etzioni & Etzioni, 2017; Greene, 2016; Hagendorff, 2020; Millar, 2014a, 2014b; Russell, 2015; Yu et al., 2018). It typically come expressed in variant examples of the 'trolly problem' – under what kinds of circumstances a life-and-death decision would or would not be permissible by an AV. The Trolley problem AV variant is used to conceptualize the relevant ethical issues. The car is unable to brake in time and is forced to choose between continuing in its path and hitting a pedestrian or swerving into oncoming traffic in an opposite lane (Bonnefon et al., 2016). An alternative variant is the car being unable to brake in time to avoid a crash just before entering a tunnel. It has two choices, in this variant, between hitting and killing a child or swerving into the wall on either side of the tunnel and killing the passenger

[13] "Driverless" or "self-driving" vehicles are used as alternative terms. These will be used interchangeably.
John Harris points out that "… so-called autonomous vehicles are incapable of deciding; they will merely do some programmed selecting between alternatives. Better surely to call them "driverless vehicles" rather than "autonomous vehicles." (Harris, J. p2, 2020).
Without the learning system in AVs, they are driverless vehicles. However, with the learning system, driverless vehicles would learn and update traffic patterns, including other human road users' behavior on the road daily. Thus, driverless vehicles are autonomous vehicles in ways that they could "make decisions" based on the initial encoded rules and daily updated relevant information. For example, they *decide* and *choose* to take another route for the same destination today compared to yesterday.
Throughout my dissertation, it is assumed that there is no learning system in "driverless vehicles". But for the purpose of simplified writing, "autonomous vehicle" and it's abbreviation, AV is used as an alternative term.

(Millar, 2014b). These thought experiments suggest that driverless cars and other AI-equipped machines[14] that make decisions on their own seem to need ethical guidance.

Unlike human-driven cars, such decision-making is predetermined for the AV, the decision being based on a deliberation on choosing between two unavoidable harms. Such predetermined decisions also have moral implications because the harm can be intended or unintended. Millar (2014a) states that choosing how AV should react when faced with an unavoidable crash scenario is one of ethical challenges in AV. Lin (2013) mentions that conflicting values as one of other factors come into play in ethics. Lin (2013) also claims that even if we do not like thinking about uncomfortable and difficult choices, such as "is it better to save an adult or child?" or "what about saving two adults versus one child?", programmers may need to do just that. Programmers still need to instruct an AV on how to act for foreseeable scenarios and implement principles for unforeseen scenarios. It means that programmers need to confront this decision, even if we human drivers never have to in the real world. Human drivers may be forgiven for making an instinctive decision in dilemmas, but programmers and designers of AVs cannot be excused for this, since they do have the time to make such an ethical decision and bear more responsibility for bad outcomes.

Within philosophy, there has been an ongoing discussion on whether AVs should be equipped with 'ethics settings' that would help to determine how they should react to the possible accident scenarios where people's lives and safety are at stake (Gordon & Nyholm, 2021; Millar, 2014b; Nyholm, 2018a, 2018b; Nyholm & Smids, 2020). Millar (2014b) claims that autonomous and smart machines such as driverless cars do not necessarily have an ability to make ethical choices autonomously. According to Sparrow and Howard (2017), for consequentialists, driverless vehicles should be introduced when they are safer than the average driver and reduce the road toll. However, Kantian ethics emphasizes our obligations to individuals understood as ends in themselves and concludes that we should not encourage the introduction of driverless vehicles until they are safer than most drivers. There is no consensus about the ethical decisions made by humans. Nonetheless, Sparrow and Howard (2017) express their opinion that philosophical debate is unlikely to delay the introduction of driverless vehicles. Instead, it may create a mixed traffic environment for a long time. They state that mixing driverless vehicles and human drivers poses challenges for humans including cyclists and pedestrians because such traffic participants do not anticipate the actions of a driverless vehicle. The presence of human drivers on the roads will also pose a special challenge to

[14] Machines guided by AI or smart machines (Etzioni & Etzioni, 2017).

driverless vehicles because the likelihood that human drivers will quickly learn what the programmed behavior of their driverless counterparts is and how to take advantage of it; "if autonomous vehicles are designed to prioritize the safety of their occupants as well as of other persons, then human beings may learn to cut in front of autonomous vehicles – or not to give way to them – on the assumption that the autonomous vehicle will give way every time in order to avoid an accident." (Sparrow & Howard, 2017, pp. 210 - 211). This is another ethical issue concerning human-AI interactions.

Ethics in human-AI interactions incorporate ethical considerations into artificial agents which are designed to influence human behaviors. How AI can act in expressing its ethical judgement is the current focus of ethical human-AI interaction research (Floridi et al., 2018). Floridi et al. (2018) suggest considering how to interact with AI in a more substantive and practical way for human dignity and flourishing. How human's decisions and choices could rely on AI solutions to be implemented and facilitated without impeding human dignity and flourishing. Yet, there is risk that AI may lead to unintended changes in human behaviors arising in response to accommodating the routines that make people's lives easier. Bosch and Bronkhorst (2018) address how humans and AI-systems should cooperate to achieve better military decision making. Bosch and Bronkhorst (2018) claim that AI systems should be functioning as intelligent team player to boost human-machine performance, but at lower levels of collaboration. Yu et al. (2018) mention that the incorporation of ethical considerations into AI systems will influence human-AI interaction. That is, by knowing that AI decisions follow ethical principles, some people may adapt their behaviors to take advantage of this and render the AI systems unable to achieve their design objectives. When attempting to influence people's behaviors, the authors claim that the benefits and risks should be distributed fairly among the users regardless of their personal backgrounds such as race, gender, and religion. March (2021) reviews the literature on experiments about the analysis of strategic interactions between human and computer players. Among the findings, there are a few impressive lessons. When human subjects interact with computer players, they adapt to computer players and often behave more rationally and more selfishly. Human subjects also adopt to computer player's type and may even learn to purposefully exploit them. More sophisticated computer players may outperform and exploit human subjects too. However, computer players may also be used to facilitate cooperation with human subjects in social dilemmas and enhance efficiency.

3. The Challenges of Moral Algorithms in AV

The persisting problem of AV ethics is how to embed heterogeneous moral preferences in the context of protecting different traffic participants to guide the decisions of an AV in dilemma situations. The difficulty of designing moral and ethical AV is discussed in this section.

3.1 The Ethical Dilemma and Its Importance

The literature at present emphasizes the importance of solving the multiple ethical preferences dilemma (Bonnefon, 2016; Bonnefon et al., 2016; Bonnefon et al., 2019; Whittlestone et al., 2019), while also recognizing the challenge of designing autonomous machines with our moral values embedded (Coeckelbergh, 2019; Lin, 2013; Lin et al., 2017; Wallach & Allen, 2009; Wallach et al., 2010). A variety of ethical and legal frameworks have been proposed as a basis for designing ethical AI and autonomous systems. These frameworks share the common characteristic that decision-making must consider multiple potentially conflicting factors (Vamplew et al., 2018). Several frameworks explore constraints to guide the autonomous system into acceptable behavior when it finds itself facing a dilemma (Nallur, 2020; Tolmeijer et al., 2021). Whether an AV should be able to choose what or who to crash into in unavoidable situations is a kind of examples that are often used to the relevance of embedding ethics into an AV to show its ethical behaviors (Bonnemains et al., 2018). Bonnemains et al. (2018) state that making such a choice implies regretting for another outcome, so in order to deal with ethical dilemma situations, an AV needs to have sufficient situational awareness.

Wallach and Allen (2009) define moral dilemmas as that "require some deliberation arise periodically in the form of conflicting voices in one's head." (p. 181). Whitbeck states that "ethical or moral problems are often represented as conflicts between (usually two) opposing sides or opposing principles, but they are often better understood as problems in which there are multiple (ethical) constraints which may or may not turn out to be satisfiable simultaneously." (Wallach & Allen, 2009, p. 75).

The Moral Machine is an online experimental website for collecting large-scale data on how citizens around the world would want AVs to solve moral dilemmas, life-threatening dilemmas in the context of unavoidable accidents (Awad et al., 2018). It displays a series of situations based on the trolley dilemma that allow the complexity of autonomous car programming in case of unavoidable accident to be comprehended. Although there is no explicit definition of moral and ethical dilemmas in this experiment, its concept can be inferred from the following.

> Think of an autonomous vehicle that is about to crash, and cannot find a trajectory that would save everyone. Should it swerve onto one jaywalking teenager to spare its three elderly passengers? Even in the more common instances in which harm is not inevitable, but just possible, autonomous vehicles will need to decide how to divide up the risk of harm between the different stakeholders on the road. Car manufacturers and policymakers are currently struggling with these moral dilemmas. … test whether people prefer to spare many lives rather than few; or whether people prefer to spare the young rather than the elderly; or whether people prefer to spare pedestrians who cross legally, rather than pedestrians who jaywalk… (Awad et al., 2018, p. 59).

According to Kirkpatrick (2015), ethical dilemmas refer to situations in which any available choice leads to infringing some accepted ethical principle and yet a decision has to be made (Yu et al., 2018). The German Ethics Commission on Automated and Connected Driving describes dilemma as "a situation in which an automated vehicle has to "decide" which of two evils, between which there can be no trade-off, it necessarily has to perform." (Luetge, 2017, p. 551). Furthermore, it describes genuine dilemmatic decisions,

> such as a decision between one human life and another, depend on the actual specific situation, incorporating "unpredictable" behavior by parties affected. They cannot thus be clearly standardized, nor can they be programmed such that they are ethically unquestionable. Technological systems must be designed to avoid accidents. However, they cannot be standardized to a complex or intuitive assessment of the impacts of an accident in such a way that they can replace or anticipate the decision of a responsible driver with the moral capacity to make correct judgements (Luetge, 2017, p. 552).

Whittlestone et al. (2019) describe moral trade-off as a situation where two values or goals conflict, then it is not possible to get more of one without sacrificing another. According to Whittlestone et al. (2019), tensions or conflicts inevitably arise as we try to implement principles in practice, choosing between important values or goals, where it appears necessary to give up one in order to realize the other.

A dilemma and an ethical dilemma in the AV ethics are defined thereafter as follows:

> A *dilemma* is the situation where a decision is made between two conflicting options in unavoidable situations. A decision leads to certain harm to either option.
>
> In AV settings, there is an *ethical dilemma* in the context of protecting different stakeholders, such as pedestrians and passengers of AV, from possible harm. The dilemma arises because it is not always possible to keep everyone equally safe, and an AV must choose which stakeholder's safety is to be prioritized. In the ethics of AV, it is not about a choice of who to kill, but *who to protect*. The practical problem is the question of how AV should be programmed to decide what to do in the dilemmatic situations. The opportunity to program these

> decisions in advance means that we can make deliberate choices. Time and urgency are not factors to be considered in the generic dilemma of AV, because the engineers of AV's system do have the time to decide what AV will value and program it in advance.

Vamplew et al. (2018) claim that even if current AVs involved in an unavoidable accident are not yet explicitly reasoning in ethical dilemmas, they do regularly make decisions which carry an implied trade-off between the safety of the driver, passengers, and other road users.
Whittlestone et al. (2019) emphasize the importance of conflicts as an important way of bridging the gap between abstract ethical principles and specific cases, and suggest that the field of AI ethics should focus more on identifying and attempting to resolve conflicts that arise when we apply them to specific cases. The current lists of principles for AI ethics are too broad and high-level to guide ethics in practice. They claim that most tensions cannot be resolved straightforwardly. Identifying tensions needs to consider how different values and goals might come into practice in concrete cases. By articulating tensions and finding ways to resolve them in specific scenarios, we can develop frameworks and guidelines for AI ethics that are rigorous and relevant.

3.2 The Heterogeneity of Moral Preferences

Page (2012) show that considerable heterogeneity can exist between people in their preferences for the principles.

The specific question of how AVs should be programmed to handle an ethical dilemma in the context of protecting different traffic participants and AV passengers has been discussed directly in, for example, (Bonnefon, 2016; Bonnefon et al., 2016; Bonnefon et al., 2019; Coeckelbergh, 2019; Dignum, 2018; Etzioni & Etzioni, 2017; Greene, 2016; Hagendorff, 2020; Russell, 2015; Yu et al., 2018).

Different stakeholders reveal different moral preferences for the choice of AV in protecting them from harm. In a series of survey experiment, Bonnefon (2016); Bonnefon et al. (2016); Bonnefon et al. (2019) investigated what moral algorithms people are willing to accept as car owners. The studies show paradoxical phenomena, such as people prefer their own AV to protect the passenger's life, but they would like others to buy 'utilitarian' AVs that maximize the number of people's life at the expense of sacrificing the passenger. Most respondents want driverless cars to make utilitarian decisions if they are not involved. However, for themselves they desired cars that will prioritize their own well-being at the cost of others.
The Moral Machine experiment (Awad et al., 2018) explored the ethics of AV by asking the public on which tradeoffs they would make during a life-threatening accident involving an AV.

It has been demonstrated that a great variation in people's attitudes exist when it comes to the question of what self-driving cars should be programmed to do in various crash dilemma scenarios (Awad et al., 2018; Gordon & Nyholm, 2021). It shows that people prefer the AV to make utilitarian sacrifices, namely prioritize saving more lives. If an AV can save more pedestrian lives by endangering its passengers, more people prefer other AVs to have this feature rather than their own AV.

Manufacturers and potential users of AV show heterogeneous moral preferences in AV experiments (Awad et al., 2018; Awad et al., 2020; Bonnefon et al., 2016; Bonnefon et al., 2019; Greene, 2016; Zhu et al., 2022). A moral gap exists between expectation and reality, which challenges manufacturers on what "ethical" types of self-driving cars to develop and produce.

Awad et al. (2020) analyze responses to variants of the trolley problem by 70,000 participants in 10 languages and 42 countries. Their cross-cultural study shows a complex pattern of universals and variations in human morality in response to when it is acceptable to sacrifice one life to save many. The study shows that there is the same qualitative ordering of sacrifice acceptability. People universally prefer Switch to in Loop, and in Loop more than in Footbridge[15]. However, there is a heterogenous moral preference in the quantitative acceptance of sacrifice in Switch, Loop, and Footbridge by different cultures. This approach is criticized in a way that self-reported preferences have been shown to deviate from actual choice behaviors. Therefore, how much the findings reflect actual choices is still an open question (Yu et al., 2018).

Zhu et al. (2022) replicate Bonnefon et al. (2016)'s experiment through an online crowdsourcing platform in China and assess people's moral preferences and expectations for AVs. The participants are most willing to purchase an AV programmed to protect passengers, followed by a utilitarian AV, and they are least willing to purchase an AV with a random algorithm. However, individuals with a pedestrian perspective prefer utilitarian AVs to

[15] "The Switch, a trolley is about to kill five workers, but can be redirected to a different track, in which case it will kill one worker. The Footbridge, a large man can be pushed in front of the trolley. The large man will die, but his body will stop the trolley before it can kill the five workers on the track. Two important differences between Switch and Footbridge: First, the death of worker in Switch is not instrumental in saving the five. It is an incidental yet foreseeable side effect of the action that saves the five. In Footbridge, the death of the large man is instrumental in saving five. His death would be a means to save them. Second, the sacrifice of the large man in Footbridge requires the use of personal force against him, whereas no personal force is exerted against anyone in Switch…. In the Loop scenario, the trolley can be redirected to a different track, where it will kill one worker whose body will stop the trolley before it can kill the five. Personal force is not required against anyone…" (Awad et al., 2017, p. 1-2).

minimize casualties, while those with a passenger perspective express a preference for AVs programmed to protect passengers. For both moral preference and expectation, participants with a pedestrian perspective had a significant preference for utilitarian AVs, compared to participants with a passenger perspective. Zhu et al. (2022) have also shown that professionals in the AV industry such as manufacturers and software developers have significantly higher perceived morality than that of the non-professional group. People in the AV industry attach moral value to AVs that prioritize passenger protection. People outside the AV industry prefer utilitarian AVs programmed to minimize casualties, although they are more likely to buy AVs that protect passengers. Both groups show a higher purchase intention for passenger-protecting AVs. Their results show that professionals tend to have a passenger perspective. Participants believe that manufacturers and software developers have more responsibility for determining the moral norms for an AV than they would for a normal car (Zhao et al., 2016). Their findings also confirm paradoxical moral preferences in AV (Bonnefon et al., 2016). Zhu et al. (2022) claim that for this reason, it seems unlikely that AVs will be programmed with a one-size-fits-all algorithm. They suggest considering how to integrate the interests of different parties involved with autonomous-driving roads.

3.3 The default AV's problems

The ethical decisions that are not prescribed by the law are left to be made by individuals or their cars. It has already been discussed that seeking to program these AVs to be able to make ethical decisions on their own is a very difficult task. People tend not to engage in such decision-making if they must make more than a few choices. Numerous studies of human behavior, ranging from retirement contributions to organ donations to consumer technology reveal that the majority of people will simply choose the default setting, even if the available options to them are straightforward and binary (Etzioni & Etzioni, 2017).

'Default' AV (DAV henceforth) is defined as a single type of risk-averse (or very 'defensive') AV that is programmed to always obey the traffic rules regardless of any circumstances on the road for safety. For example, DAV will slow down and wait for pedestrians to cross even at unmarked crosswalks or even for jaywalkers. DAV has two crucial issues. First, DAV causes moral hazard behavior from human road users. Because there is a certainty that DAV is programmed to stop always for pedestrians and other human road users in any situations on the road, some of the traffic participants may exploit this safe and predictable technology. This is what Millard-Ball (2018) expresses a concern about pedestrian supremacy. Suppose there is only a DAV that always protects pedestrians on any occasion for

sure. If pedestrians know about this aspect of AV's technology, they are free to cross the road anytime and it leads to pedestrian supremacy. This problem also causes transportation inefficiency in mixed traffic.

Lin (2013) raises concern about the abuse and misuse of AV. If the cars drive too conservatively, they may become a road hazard or trigger road-rage in human drivers with less patience. If the crash-avoidance system of an AV is known, other drivers may be tempted to "game" it, e.g., by cutting in front of it, knowing that AV will slow down or swerve to avoid an accident.

DAV is unable to address heterogeneous moral preferences in the context of protecting different participants in mixed traffic. Previous studies show that people prefer other people's AV to protect pedestrians even at the expense of sacrificing the passengers. But they prefer their own AV to prioritize protecting passengers. It is impossible to bridge this moral gap between expectation and reality with DAV. These are what Zhu et al. (2022) also point out that programming AVs with a one-size-fits-all algorithm would be unlikely.

4. Two types of reasoners in an AV decision process

As discussed in the previous Section 3, Bonnefon et al. (2016) and Zhu et al. (2022) mention about different desirable types of AVs. It points to two different types of reasoners that can be implemented in AVs. In this section, two types of reasoners, one that prioritizes the protection of its passengers, and the other that prioritizes the protection of other traffic participants outside of the vehicle are defined.

4.1 A New Two Types of Reasoners[16]

How should an AV reason to make decisions in an environment where people are committed to taking unethical actions? So far in the literature, it is always assumed that humans are superior reasoners to the AV and that they behave ethically. Deviated from that assumption and assume that there are two types of human participants (or stakeholders) in the AV decision process: people in the car and people out of the car. The 'immoral' intention of people in the car and out of the car causes the ethical dilemma in AV.

Consider, for example, a car hijacking scenario where car hijackers attempt to steal the vehicle after forcing it to stop. In a car hijacking scenario, protecting the AV passengers from outsiders should be prioritized. On the other hand, we can consider a terrorist scenario in which

[16] I am indebted to Prof. Marija Slavkovik for this idea and helpful discussions about the relevant issues during the research visit at the Department of Information Science and Media Studies, University of Bergen in Norway.

a terrorist wants to use the vehicle to cause harm to pedestrians. In a terrorist scenario, protecting pedestrians as outsiders of an AV should be the concern. Our approach considers other human road users, such as cyclists and motorcyclists as outsiders of AV.

This new approach provides the reason for the existence of two different types of AV that also solve the ethical problems and issues with DAV. Arguing that studying ethical dilemmas that are caused by a malfunction of the vehicle is unrealistic. One argument against programming for possible malfunction is the expectation that the autonomous vehicle engineers would seek to put control of the vehicle in the hands of a human operator as soon as the malfunction of the AV is detected. If an AV malfunction does cause a moral dilemma, that AV is likely to be recalled. Regulations are currently being developed to define machines[17] that can cause loss of life as high-risk AI systems in Europe (Commission, 2021; European et al., 2020). A high-risk AI system that causes a serious accident would be considered to be malfunctioning and would be recalled or redesigned after the first accident. Millar (2014a) states that a malfunctioning smart car is not 'autonomous' in the way that people are, because there appear to be no moral injunctions against implementing extensive constraints on a smart car's actions. The car is a malfunctioning tool that should be dealt with accordingly. Therefore, the problem of deciding whether to prioritize the life of pedestrians or passengers is not one that would be left to the AV (Millar, 2014a). However, this does not mean that the AV does not need to solve dilemmas related to harm inflicted on traffic participants.

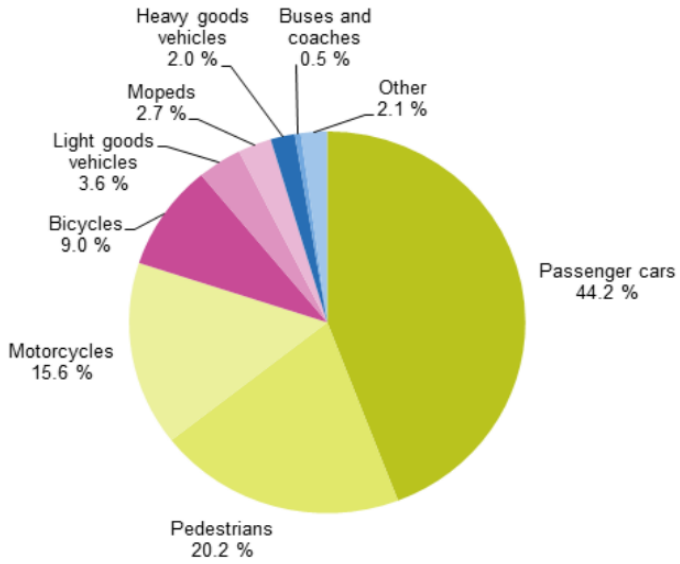
4.2 Insider and Outsiders Protection Priority AV

Most analyses of AV ethics consider only pedestrians in comparison to the passengers of the AV. In most previous studies and typical experiments on moral decisions of AV, people categorize AV passengers versus pedestrians as the reasoner in the AV decision process. Participants face an unavoidable crash and must decide whether to save the pedestrians on the road or the passengers in the AV. The passenger protection AV is described as the opposite concept of utilitarian AV, and it is not the feature of utilitarian AV. Here, it is claimed that comparison of self-protective (protecting the passenger) and utilitarian AVs is incompatible. The passenger protection AV could be a utilitarian AV and it depends on the number of passengers. For example, if there are four passengers in an AV and two pedestrians on the road, passenger protection AV would work as equivalent to utilitarian AV in unavoidable situations.

[17] Machines that can be involved in any serious incidents that directly or indirectly leads, might have led, or might lead to, among else, the death of a person or serious damage to a person's health, to property or the environment.

There are also other types of participants than passenger and pedestrian, such as cyclists, motorcyclists, and human drivers on the road[18].

A new approach to categorize the reasoner in AV decision process is proposed as insiders and outsiders of AV. The passengers of an AV are regarded as insiders of the AV and the outsiders of the AV are composed of pedestrians, cyclists, and motorcyclists[19]. These new two types of reasoners cover a broad range of human road users and embeds their different moral preferences in mixed traffic. The occupants of the passing AV are passengers of that AV. However, they are the outsiders of my own AV. The occupants of my own AV are the insiders of my own AV, but they are outsiders of the passing AV. This enables car manufacturers to develop two different types of AV – insiders vs. outsiders protection priority, without raising much public outrage for the only passenger protection AV. Two different types of AV resolve ethical issues with DAV. A new approach also contributes to the freedom from an ethical question such as whether two passengers should sacrifice to save one pedestrian in each crash. Because the fatality ratio of insiders to outsiders of AV is (approximately) one to one from Graph 1.



Graph 1 Road accident fatalities by category of vehicles, EU, 2019[20]

[18] In some countries, there are even more cyclists and motorcyclists than pedestrians. For example, there are more cyclists than pedestrians in most cities in the Netherlands. In Italy, there are more motorcyclists than pedestrians in some locals. Therefore, it is reasonable to consider cyclists and motorcyclists in the analyses of AV ethics.

[19] Human drivers could be also regarded as outsiders of AV.

[20] Source: File:Road accident fatalities by category of vehicles, EU, 2020 (%).png - Statistics Explained (europa.eu)

Consider the road accident fatalities by category of vehicles in 2019 in the European Union. Of these, about 44.2% were car passengers (drivers included), and 20.2% were pedestrians. In other words, about two passengers died for one pedestrian. However, categorizing car passengers as insiders of cars and pedestrians, cyclists, and motorcyclists as outsiders of cars shows that about one insider of car dies for one outsider of car, since 44.8% were outsiders. Hansson et al. (2021, p. 1396) mention that if the coordination between automatized vehicles is efficient, most accidents will probably result from collisions with cars driven by humans and with unprotected travelers such as pedestrians, cyclists, and motorcyclists.

5. The Strategic Interaction between Pedestrian and Two types of AV

Game theory is one of possible ways towards developing a general ethical decision-making framework for AI. Standard game theory is used to solve tensions between different rational agents that maximize their own payoffs, regardless of the consequences on the other agents in the economics literature. However, many ethical principles are based on social emotions, such as compassion and empathy in common. Humans apply different moral values for themselves and to others. A standard game theoretic approach could be adjusted to take account of the ethics and applied to resolve tensions between the conflicted ethical preferences or different ethical principles.

5.1 Related work

It was Braithwaite (1955) who first showed how game theory could be used to arrive at moral choices and ethical decisions.

Conitzer et al. (2018) propose two ways towards developing a general ethical decision-making framework for AI based on game theory. The authors claim that using game theory is one approach to providing guidance on moral behavior and arriving at general purpose procedures for making moral decisions for AI. Moral concepts such as selfishness, loyalty, trustworthiness, and fairness influence the actions that people choose to take. Rather than the simplistic game-theoretic solution which fails to account for ethical considerations, Conitzer et al. (2018) introduce moral solution concepts in game theory with an example of the trust game. It is the analysis of the game that must be intertwined with the assessment of whether an agent morally should pursue another agent's well-being. Therefore, in such situations, game theory is potentially a good fit for abstractly representing moral dilemmas. Conitzer et al. (2018) also suggest the extensive form (a generalization of game trees) as a foundation scheme to represent dilemmas. As the current extensive form does not account for protected values in which an action can be treated as unethical regardless of its consequence, they propose to extend the

extensive form representation with passive actions for agents to select to be ethical. In particular, they suggest that game theoretic analysis of ethics is used as a feature to train machine learning approaches, while machine learning helps game theory identify ethical aspects which are overlooked (Yu et al., 2018).

Millard-Ball (2018) apply game theory to analyze pedestrians' yielding behavior at crosswalks with human-driven vehicles and AV. The author uses a traditional payoff matrix to address the interaction between pedestrians and regular human-driven cars that follows a chicken game. The best response to each other is for the car to go and the pedestrian to stop avoid injury (or the car to stop and the pedestrian to go and cross the road). Upon analyzing the interaction between pedestrian and AV, Millard-Ball (2018) assumes that there are sufficiently few human drivers to prevent the equilibrium in the crosswalk chicken game. In addition, AV will be risk averse. Because it is programmed to obey the traffic rules, including slowing down in a risky situation and waiting for pedestrians to cross even at unmarked crosswalks. Given these circumstances, Millard-Ball (2018) claims that pedestrians and other human road users, including drivers in human-driven vehicles and cyclists possibly could exploit predictable, safe, and risk averse behavior of AVs. It is implied that the more cautiously an AV will behave, the more road users act recklessly such as a pedestrian stepping onto the road and human driver ignoring a red light. Because pedestrians and other human road users know and believe that AV will obey the rules and stop at crosswalks. AVs are unable to win the game of chicken. Therefore, *Go* becomes the pedestrian's dominant response and the AV's best response to the pedestrian is to stop.[21] This is great for safety but incentivizes people to behave carelessly, which can easily cause a lot more traffic on the road.

Potential owners of an AV may not want to purchase such an AV that regards careless behavior of pedestrians as unaccompanied children and drives slowly down on a street. This AVs' strategic disadvantage aspect may hinder the adoption of AV in urban areas and lead to pedestrian supremacy. Millard-Ball (2018) also suggests that changing laws to reduce pedestrian priority would be one of regulatory response to solve these issues. Introducing insiders' protection priority AV as different type of AV other than outsiders' protection priority AV could be one of solutions to resolve the issues that the author raises.

de Melo et al. (2019) show that people want to program their AVs to be more cooperative than they would if driving themselves. In their experiments, when programming an autonomous

[21] Millard-Ball (2018) regards it as 'pedestrian supremacy'. The author does not use the payoff matrix to show it. In Section 5.3, the default AV (DAV) is assumed to demonstrate Millard-Ball's concern of pedestrian supremacy as a pure Nash Equilibrium: (Pedestrian, DAV) = (Go, Stop) in the static game model.

machine, people are asked to think about the situation and decide ahead of time. In the context of testing game theory, it introduces an opportunity to make a precommitment by programming the machine ahead of time. When people are asked to report their decisions ahead of time, they tend to behave more fairly than when making the decision in real time. People are more inclined to cooperate when programming an agent than when engaging in real-time direct interaction. Programming vehicles ahead of time from the owner's perspective leads to increased cooperation. That is, people are more willing to sacrifice individual for collective interest than if driving the vehicle directly. This effect prevails even if participants are allowed to reprogram their machines independently of whether others behaved cooperatively or competitively. The work of de Melo et al. (2019) indicates that autonomous machines, including AV have the potential to shape how the social dilemmas are solved with people being more cooperative, and thus stakeholders have an opportunity to promote a more cooperative society.

Di et al. (2020) use a hierarchical game-theoretic approach to examine how the liability rule should evolve with the existence of self-driving car. Different game models have been adopted with respect to the strategic interactions among the lawmaker, the AV manufacturer, AV, and human driver (human-driven vehicle). Between human drivers, they apply a unified game, including a Nash game in a way that human drivers simultaneously choose 'care levels' or 'the level of precaution' to minimize their own cost. Di et al. (2020) apply Stackelberg game model (leader-follower game) between the AV manufacturer and human-driven vehicle. Human drivers are strategic players because they can choose care levels. Unlike human drivers, AVs are not strategic players because their care levels are predetermined by the manufacturer. Therefore, the AV manufacturer and human driver play a Stackelberg game as the leader in which manufacturer stipulates a care level for its AV and human driver select their care levels. A Stackelberg game is also applied between the lawmaker and other users. The lawmaker is the leader, because they make a high-level decision on an optimal driver liability rule for human drivers and a product liability rule for AV manufacturers to minimize the social cost. One AV manufacturer is the middle level player who determines driving settings. A human driver is the lower-level player, as the follower who selects the level of care to maximize their own utility. Their approach shows the possibility of human drivers' moral hazard with AV. Di et al. (2020) claim that an optimal liability rule design is important to prevent the moral hazard of human drivers and improve social welfare because if human drivers realize their road environment has become safer, they could develop moral hazard.

5.2 Structure of Game

A strategic game is a model of interacting decision-makers as players. Each player has a set of possible actions. The model captures interaction between the players by allowing each player to be affected by all players' actions, not just her own. Each player has preferences about the action profile – the list of all the players' actions.

A strategic game is defined as follows.

DEFINITION 1.1 (*Strategic game with cardinal preferences*) A **strategic game** (with cardinal preferences) consists of

- a set of **players**
- for each player, a set of **actions**
- for each player, **preferences** over the set of action profiles.

The strategic interaction between pedestrian and autonomous vehicles is regarded as a case of non-programmed human and programmed vehicles as below in Table 2.

		Vehicle	
		Non-programmed (Driver vehicle)	Programmed (Driverless vehicle)
Pedestrian[22]	Non-programed	Standard game	This case
	Programed		

Table 2 Pedestrian as "non-programed" agent and AV as "programmed" vehicle.

5.2.1 Assumptions

The scenario in which a pedestrian is illegally crossing the street intercepting the path of an AV is considered in this section.

The following assumptions are required to translate ethical considerations into computable concepts.

[22] In game theory, it is assumed that the players (agents) are *rational*. In this context, people may argue that rational agents such as pedestrians are regarded as *programmed* agents. But I claim that there is a difference between rational and programmed agents. For example, "Sophia", a robot granted citizenship by Saudi Arabia in 2017 would be a programmed agent, and the citizens are not called programmed agents but rational agents. Furthermore, when people normally take moral decision making, they practice with "bounded morality" given abstracts and broad ethical principles and rules. It is the concept of bounded rationality in ethics and the players are assumed to be *bounded rational* rather than perfectly rational. This is called *ethical heuristics* and it is a concept of following rules that are expected to increase local utility rather than analyze all the consequences of one's possible actions like perfectly rational agents do.

The players are pedestrians and two different types of AV. The two types of AV are: one that prioritizes the safety of outsiders and one that prioritizes the safety of the AV. The outsider's protection priority AV is denoted as type I (AVI henceforth), while the insider's protection priority AV is type II (AVII henceforth). Outsiders of AV are pedestrians, cyclists, and motorcyclists. It is also assumed that AVI and AVII has same optimization goals as acceptable common conclusions[23], such as carrying passengers to the desired destination, safe, fuel-efficient and travel time efficient (Nyholm & Smids, 2020).

The difference between a pedestrian and an AV is explicit and implicit ethical player/agent, respectively. AVI and AVII are implicit ethical players[24] that do not learn or encode ethics explicitly. It means they cannot choose among the different kinds of harm in a situation when their actions impact ethical issues such as safety and are constrained to unavoidable unethical outcomes such as fatalities. For example, when implicit ethical AVs are in a situation where a crash is unavoidable and when all trajectories are likely to end up in casualties, they find themselves stuck and are unable to choose among the different ethical choices. From an engineering perspective, situations in which AVs would act as explicit ethical agents seem too far-fetched to be a priority (Bonnefon et al., 2019). According to Bonnefon et al. (2019), the German Ethics Code for Automated and Connected Vehicles considers an aspect of implicit ethical AV in a way that AV ought to save lives and eliminate at least some traffic fatalities which is one of the key promises of automated driving: "The primary purpose of partly and fully automated transport systems is to improve safety for all road users…." Bonnefon et al. (2019) also state that it is far easier to treat AVs as implicit ethical agents, who are systematically programmed to minimize the absolute risk of a crash. This relates to responsibility and liability issues in AVs, because it could be regarded as shifting relative risk

[23] According to the floating conclusion as skeptical reasoning, acceptable common conclusions from conflicting propositions are justified (Horty 2002; Bonnefon 2004). This approach is validated in the context of economics in Section 2.2.

[24] This can be thought of as "weak" or "narrow" AV such that only process specific and programmed tasks. The performance of narrow AV may depend on the training data and programming, which is relevant to big data and humans. Therefore, the related ethical issues of AV are closely related to human driving and yield behaviors in traffic. Constraining the machine's actions to avoid unethical outcome is the way to educate AI system to be an implicit ethical agent. To train AI into explicit ethical agents, we need to state explicitly what action is allowed and is forbidden (Moor, 2006).

from one type of road user to another.[25] Such a shift can be also recognized as an ethical tradeoff.

Unlike implicit ethical AVs, pedestrians are explicit ethical players (or explicit moral reasoners) who make life and death decisions based on encoded ethics. They encode ethics explicitly, learn, and update ethical behavior. They can arbitrate between different kinds of harm in unavoidable situations, if any. The unique strength of human intelligence is its ability to learn and adapt to new environments and challenges. Refining and improving performance through continuing learning has been a challenge for advancing AI until the recent advances in deep learning and Big Data (Duan et al., 2019). Baker-Brunnbauer (2021) also mentions implicit and explicit ethical agents as the potential ways to educate AI systems for ethics. Forcing the machines' actions to prevent unethical outcome means to educate AI for an implicit ethical agent. Explicit ethical agents are explicitly quoted for the allowed and forbidden actions.

The players' preferences are represented by payoff functions. These payoffs are *cardinal* numbers, not ordinal. The difference in numbers indicates not only the preference order, but also the degree of preference. That is, it informs about how much better the preferred outcome is. Cardinal payoffs enable us to reflect the protection priority of AV. Given the default AV's payoffs, we attach cardinal moral value to construct protection priority of the insiders and outsiders of AV, respectively. Thus, the preferences of an AV reflect the players' protection priority. This features two different types of AV in the ethical context. When we deal with decisions involving uncertainty and ethics, the cardinal measure is required (Bowles, 2021).

Consider the risk attitude in the payoffs of pedestrian and two types of AV. The default AV (DAV) will be very risk averse such that it slows down and yields in a way that waiting for pedestrians to cross even at midblock and unmarked crosswalks. This is what Millard-Ball (2018) mention. Di et al. (2020) also state that human drivers expect and think an AV will be risk averse, therefore the human drivers will exploit its technology by careless behavior. Human drivers will take more risks and develop moral hazard if they think their road environment has become safer. In addition, propensity to take risk for pedestrian to AVI (outsiders protection priority) and AVII (insiders protection priority) is risk-neutral and risk-averse, respectively. On the contrary, the propensity to take risk for AVI and AVII to pedestrian is risk-averse and risk-neutral as shown in Table 3 as below.

[25] For example, it could be a shift in relative risk from passengers of AV to pedestrians or vice versa. In our approach, shifting relative risk is from insiders of AV to outsiders AV or vice versa.

	AVI	AVII
Pedestrian	Risk-neutral / Risk-averse	Risk-averse / Risk-neutral

Table 3 The players' risk attitude

A set of actions to take is *Go* and *Stop* for both players. Both pedestrians and two types of AVs are rational decision-makers that choose the best action according to their preferences, among all the available actions to them.

Time is absent from the games in Section 5.3, 5.4, and 5.5. The players simultaneously choose their actions once and for all. Therefore, they are all static games. It is also reminder that AVs are assumed to not have a learning system in this setting.[26]

5.2.2 The general payoffs condition[27]

		AV	
		Go	Stop
Pedestrian	Go	a, b	c, d
	Stop	e, f	g, h

Table 4 The general payoff matrix

It is a reminder that the scenario in which a pedestrian is illegally crossing the street intercepting the path of (an) AVs is considered in this section.

A row represents pedestrian's action profile and its payoffs while a column represents AV's action profile and its payoffs. The first components of each cell – a, c, e, g represent the pedestrian's payoffs.

For a pedestrian, c is the highest payoff when AV yields for her. A potential collision is the worst outcome, and it is presented as the lowest payoff a. When the pedestrian stops and yields for an AV, she gets the same payoffs of e and g *regardless of AV's actions*. The pedestrian's payoffs condition is described as follows in (1). The pedestrian condition (1) holds for the interaction with type I (AVI – outsiders protection priority) and type II (AVII – insiders protection priority).

$$c > e = g > a \quad - \quad (1)$$

[26] As an implicit ethical player, AV is unable to learn and update encoded ethics like pedestrian. This also characterizes a static game. AVs with the learning system could be modeled as a dynamic game. A dynamic Bayesian game model in AV dilemma is discussed as further research topic in Section 8.2.

[27] I am indebted to Dr. Rune Nyrup for helpful discussions about the construction of payoffs in the context of two different types of autonomous vehicle and relevant issues during the research visit at the Leverhulme Centre for the Future of Intelligence, University of Cambridge in the United Kingdom.

However, (1)' is a pedestrian's condition interacting with the default AV (DAV). Because DAV is programmed to stop for pedestrians in any situation on the road. Therefore, there is no potential collision for this pedestrian and $c = a$.

$$c = a > e = g \quad - \ (1)'$$

For an AV, f is the highest payoff, which occurs when a pedestrian yields, while b is the lowest payoff for the potential collision with the pedestrian. When an AV stops and yields for a pedestrian, the AV gets the same payoffs of d and h *regardless of the pedestrian's actions*. The AV's condition is described as follows in (2). It holds for both DAV and AVI's condition.

$$f > d = h > b \quad - \ (2)$$

However, the following (2)' is AVII's condition. Because although AVII still considers pedestrians, it prioritizes the safety of the passenger as insiders of vehicle. Therefore, *Go* is the dominant strategy for AVII: f is still the highest payoff but d and h becomes the lowest payoffs.

$$f > b > d = h \quad - \ (2)'$$

Two things are remarked. First, if the outsider protection priority AV's payoff is represented by $d > h$, the cost of introducing AVI as a different type of AV than DAV is little, while its benefits are great; Pareto efficient outcome. Because the difference in preference order between DAV and AVI is only the inequality between the payoffs d and h. That is, DAV's payoff is represented by $d = h$ and AVI's payoff is represented by $d \geq h$. It indicates that AVI is weakly preferred to stop and yield for a pedestrian compared to when both players simultaneously stop to each other. For simplicity, it will be just considered $d = h$ thereafter. The cost of introducing AVI as a different type of AVII is also small and discussed in Section 7.2. Second, the cost of potential collision, including the medical expenses, liability, and financial costs for both the pedestrian and the car manufacturer are presented as a negative payoff such as $a, b < 0$. There is a potential collision with the existence of different types of AV. However, it is assumed that there is no potential collision for a pedestrian with DAV, so a is a non-negative payoff in this case.

5.3 No Chicken game for default AV

It is assumed that the players' payoffs are common knowledge. This means that the information is *complete* in a way that a pedestrian and a default AV (DAV) know each other's

preferences. Table 5 illustrates no chicken game in the formal representation of a simultaneous move game of complete information, called a static Nash game.

		DAV	
		Go	Stop
Pedestrian	Go	c, b	c, d
	Stop	e, f	g, h

Table 5 No chicken game

A row represents a pedestrian's action profile and its payoffs while a column represents DAV's action profile and its payoffs. The payoff functions represent the players' preferences. These payoffs have a *cardinal* significance. It is a reminder that there is no potential collision for a pedestrian with a DAV. Because there is a certainty that DAV is programmed to stop always for a pedestrian in any situations on the road, its payoffs satisfy $f > d = h > b$. A pedestrian knows about this DAV's safe and predictable technology. She is confident about her 'trespass' on the road. Her payoffs satisfy conditions $a = c$ and $c > e = g$. The pedestrian only chooses to *Go*, so it is the pedestrian's dominant strategy[28]. Compared to the pedestrian's strategy, the DAV is likely to respond very differently because it is fundamentally unable to win the game of chicken. A DAV is likely to obey the rules of the road in safety-critical situations, far more than pedestrians, including other human road users. Given the pedestrian's dominant strategy to *Go*, the DAV will choose to *Stop*. That is the Nash equilibrium[29]. The notion of Nash equilibrium for a strategic game models a steady state in which each player's beliefs about the other players' actions are correct, and each player acts optimally, given her beliefs. It is generalized and claimed that pedestrians exploit DAV's safe and predictable technology to stop for pedestrians all the time[30]. This result shows Millard-Ball (2018)'s concern about pedestrian supremacy.

[28] Another way to interpret it without $c = a > e = g$ is that pedestrian regards $p = 1$ in mixed traffic. P is the portion of AVI, so all AVs are outsiders (pedestrian) protection priority. AVI stop always for the outsiders of vehicle, and no potential collision for pedestrians. Therefore, pedestrians always choose to *Go* and it becomes her dominant strategy.

[29] There are two Nash equilibria – either pedestrian goes and DAV stops, or DAV goes and pedestrian stops. Among them, the only Nash equilibrium is (Pedestrian, DAV) = (Go, Stop) in this case. This is not Pareto efficiency outcome in a mixed traffic.

[30] This bad behavior of human road users is called a *moral hazard*.

5.4 A strategic game with Type I and Type II AV

Suppose there are two types of AV in terms of protection priority of outsiders and insiders of AV[31], and the dilemma scenario in which a pedestrian is crossing the street intercepting the path of an AV. In this section, AVI and AVII are introduced, and illustrating the strategic interactions with a pedestrian separately in the formal representation of a simultaneous move game of *complete* information. The players' payoffs are still common knowledge, so a pedestrian, an AVI, and an AVII all know each other's preferences.

It is only considered different types of AV's risk attitudes for a pedestrian and present it in the payoff matrix for the analysis. A pedestrian's preference is assumed to be consistent to both an AVI and an AVII in the way that she gets the worst payoff in the potential collision but when each type of an AV yields for pedestrian, she gets the greatest payoff.

5.4.1 Pedestrian and AVI (outsiders protection priority)

In Table 4, the risk neutral pedestrian's payoff satisfies $c > e = g > a$ such that she gets the highest payoff c when the AVI yields for her. A potential collision is the worst outcome for her, so a is the lowest payoff. For a risk averse AVI, f is the highest payoff in a situation when pedestrian yields for it. A potential collision with the pedestrian is also the worst outcome for the AVI hence b is the lowest payoff: $f > d = h > b$. A pedestrian wishes to go and cross as soon as possible and prefers that the AVI stops and yields. However, she knows that although the AVI protects her as an 'outsider' of the AV, there is still chance to get hit by an AVI when she does not obey the traffic rules because there are cyclists and motorcyclists as other members of outsider than pedestrians themselves. The risk of potential collision exists, and the fatality would be distributed among them[32]. This means that the AVI's protection priority to the pedestrian needs not to be a perfect no harm guarantee for her.

There are two Nash equilibria – the pedestrian goes and the AVI stops, and the AVI goes and the pedestrian stops. This is the best scenario[33] in mixed traffic. It shows that introducing the outsider's protection priority AV resolves a possible inefficient transportation problem and the human road users' moral hazard behavior with DAV in mixed traffics.

[31] See Section 4.2.

[32] The fair distribution of fatality and its relevant liability is another issue to be discussed.

[33] The Pareto efficiency outcome.

5.4.2 Pedestrian and AVII (insiders protection priority)

The pedestrian is a risk averse player in the interacting with the AVII. The pedestrian's payoff is the same as with the AVI and satisfies $c > e = g > a$ such that she gets the highest payoff when the AVII yields for her, see Table 4. A potential collision is the worst situation for her, so a is the lowest payoff. On the contrary, the AVII is a risk neutral player for pedestrians. The AVII still "cares" for the pedestrian, but it prioritizes the safety of the passengers as insiders of the vehicle. This means that when the AVII is in ethical dilemma situations, it would keep protecting the vehicle's insiders. That is, f is still the highest payoff in a situation when the pedestrian yields for it. However, stopping the vehicle is the worst outcome for the AVII. Therefore d and h becomes the lowest payoffs such that $f > b > d = h$, see Table 2. This implies that *Go* is the dominant strategy for the AVII and given AVII's dominant strategy to *Go*, the pedestrian always chooses to *Stop*, which is the Nash equilibrium[34]. This result shows the AVII supremacy. It means that there will be no pedestrians and other members outside of the AVII on roads except AVII. It is an inefficient transportation situation in mixed traffic.

5.5 A Static Bayesian Game with AVI and AVII

Conitzer et al. (2018) suggest the game theoretical concept with more players and/or *imperfect* information as the possible extensions of their moral solution concepts and capture ethical concerns in game theory. In a game of imperfect information, at some move in the game the player with the move does not know the complete history of the game. Recall that the players simultaneously choose their actions *once* and for all in our analysis of games. Therefore, a game of *incomplete* information is introduced as a game of *imperfect* information, whereby incomplete information I mean that at least one player is uncertain about the other player's payoff function.

Now, there are two types of AV in terms of protection priority of outsiders and insiders of AV, and the dilemma scenario in which a pedestrian is crossing the street intercepting the path of an AV. The normal-form representation of a simultaneous move game of incomplete information, also called a static Bayesian game is developed and a Bayesian Nash equilibrium is defined in such a game in this section. In a game of incomplete information, at least one player is uncertain about another player's payoff function.

[34] There are two possible Nash equilibria – either pedestrian goes and AVII stops or AVII goes and pedestrian stops. Among them, the only Nash equilibrium is (Pedestrian, AVII) = (Stop, Go) in this case. This is not *Pareto* efficiency outcome in a mixed traffic compared with AVII.

A Bayesian game is defined as follows.

DEFINITION 1.2 (*Bayesian game with incomplete information*) A **Bayesian game** consists of

- A set of **players**
- A set of **states**

For a pedestrian,

- A set of **actions**
- A **belief** about the states
- A **Bernoulli payoff function**

5.5.1 Assumptions

The players are pedestrians and two different types of autonomous vehicles, an AVI and an AVII. A set of available actions for them to take are *Go* and *Stop*. The pedestrian, the AVI and the AVII are rational decision-makers that choose the best action according to their preferences, among all the available actions to them. Each player knows his or her own payoff function but may be uncertain about the other players' payoff functions. Given the definition of the AV's type, saying that an AV knows its own payoff is equivalent to saying that the AV knows its own type. Likewise, saying that the pedestrian may be uncertain about an AV's payoff is equivalent to saying that the pedestrian may be uncertain about which of the AV's types it is encountering. From the pedestrian's point of view, she knows that there are two types of AV on the roads. But she does not know which is the AV's type. That is, the pedestrian is not completely/perfectly informed about the AV's characteristics and some aspect of her environment relevant to her choices of action. So, to choose an action rationally, she needs to form a *belief* about the action of the AVI and the AVII. Given these beliefs and the belief about the likelihood of each type of an AV to be encountered, we calculate her expected payoff to each action. Their payoffs are called *Bernoulli* payoffs.

5.5.2 The general condition of Bayesian Nash Equilibrium

It is now assumed that from the pedestrian's point of view, an AV has two types, one whose outsiders-protection priority preference is given in the left table of Table 6, and one whose insiders-protection priority preference is given in the right table of Table 6. The pedestrian is incompletely or imperfectly informed about the types of AV that she is facing on roads. Therefore, she assigns the likelihood of p on encountering type I and $(1-p)$ on encountering type II of an AV, respectively to choose an action rationally.

Table 6 illustrates a static Bayesian game if p is the probability of encountering a type I AV and $(1-p)$ is the probability of encountering a type II AV, where $p \in [0,1]$. The expected payoffs of the pedestrian are derived from Table 6 and presented in Table 7. Recall that the AVI (type I) is the outsiders (pedestrians, cyclists, and motorcyclists) and the AVII (type II) is the insiders (passengers) protection priority AV.

Prob. p			
		AVI	
		Go	Stop
Pedestrian	Go	a, b	c, d
	Stop	e, f	g, h

Prob. $(1-p)$			
		AVII	
		Go	Stop
Pedestrian	Go	a, b	c, d'
	Stop	e, f	g, h'

Table 6 A Bayesian game payoff matrix

Each player knows his or her own payoffs. AVs know the relevant table; type I on the left and type II on the right table. However, pedestrians are unsure which type of AV is facing on the road. She does not know the relevant table; the probability she assigns to type I and type II are p and $(1-p)$, respectively.

A pedestrian has a consistent preference that satisfies $c > e = g > a$ in the interaction with both the AVI and the AVII. The pedestrian is better off when the AVs yield for her than when she yields for the AVs. However, the AVI's outsider-protection priority preference satisfies $f > d = h > b$ and the AVII's insider-protection priority preference satisfies $f > b > d' = h'$, where $a, b < 0$, see Table 6. Because the payoffs are represented with a cardinal preference, they characterize the different types of AV in the context of whom to protect "first" without loss of generality. For example, consider the preference order such as $d' = -0.2 < b = -0.1 < d = 0$. The characteristic of the AVI is represented with a payoff $d = 0$ and the AVII is characterized with a payoff $d' = -0.2$, given the same payoff of potential collision $b = -0.1$. When each type of an AV yields for a pedestrian, the outsiders protection priority AV's payoff is greater than that of the insider's protection priority AV: $d > d'$.

Given these beliefs and the pedestrian's belief about the likelihood of each type, we can now calculate her expected payoff to each of her actions. For example, given the belief about the action pair of AVI and AVII is *Go* and *Stop*, choosing *Go* yields her a payoff of a with probability p and a payoff of c with probability $(1-p)$, so that her expected payoff is $p \cdot a + (1-p) \cdot c = c - p(c-a)$. Choosing *Stop* yields her an expected payoff of $p \cdot e + (1-p) \cdot g = g + p(e-g) = g$ because $e = g$ from a pedestrian's preference condition. Similar calculations for the other combinations of actions for the two types of AV yield the expected payoffs in Table 7.

Each column of the table is a pair of actions for the two types of AV. For example, a pair of actions (Stop, Go), the first member of a pair being the action of AVI to *Stop* and the second member being the action of AVII to *Go*.

	(Go, Go)	(Go, Stop)	(Stop, Go)	(Stop, Stop)
Go	a	$c - p(c - a)$	$a + p(c - a)$	c
Stop	e	g	e	g

Table 7
The pedestrian's Bernoulli payoffs for the four possible pairs of actions of the two types of AV

For this situation, we can define a pure strategy of a *Bayesian Nash equilibrium* to be a triple of actions, one for a pedestrian and one for each type of AV, with the property that

- the action of pedestrian is optimal, given the actions of the two types of AV (and pedestrian's belief about the state)
- the action of each type of AV is optimal, given the action of pedestrian.

That is, we treat the two types of AV as separate players, and analyze the situation as a three-player strategic game in which the pedestrian's payoffs are a function of the actions of the two other players (i.e. AVI and AVII), as given in Table 7, and the payoff of each type of AV is independent of the actions of the other type and depends only on the action of the pedestrian as given in the tables in Table 6. In a Nash equilibrium, the pedestrian's action is a best response in Table 7 to the pair of actions of the two types of AV, the action of the AVI is a best response in the left table of Table 6 to the action of the pedestrian, and the action of the AVII is a best response in the right table of Table 6 to the action of the pedestrian.

Given the actions of the two types of AV are (*Go, Go*), pedestrian's action *Stop* is optimal in Table 7, because $e > a$. Given that pedestrian chooses *Stop, Go* is optimal for both AVI and AVII in Table 6. Therefore, (*Stop,* (*Go, Go*)) is a Bayesian Nash equilibrium.

Given the actions of the two types of AV are (*Stop, Stop*), *Go* is the pedestrian's optimal action in Table 7, because $c > g$. Given that the pedestrian chooses *Go, Stop* is optimal for the AVI because $d > b$, but *Go* is optimal for the AVII because $b > d'$ in Table 6. Therefore, (*Go*, (*Stop, Stop*)) is a not a Bayesian Nash equilibrium. Instead, we check whether (*Go*, (*Stop, Go*)) could be another Bayesian Nash equilibrium. Given the actions of the two types of AV are (*Stop, Go*), *Go* is optimal for pedestrian, if and only if $a + p(c - a) \geq e$ from Table 3. That is, $p \geq \frac{e-a}{c-a}$ is the condition for a Bayesian Nash equilibrium of (*Go*, (*Stop, Go*)). (*Stop*, (*Stop, Go*)) cannot be a Bayesian Nash equilibrium, because *Go* is optimal for both AVI and AVII given the pedestrian's action profile of *Stop* in Table 6. We know that given that the pedestrian chooses

Go, (*Stop*, *Go*) is the optimal action pair of the AVI and the AVII. Therefore, (*Go*, (*Go*, *Stop*)) cannot be a Bayesian Nash equilibrium. Likewise, given that the pedestrian chooses *Stop*, (*Go*, *Go*) is optimal for both the AVI and the AVII. Thus, (*Stop*, (*Go*, *Stop*)) is not a Bayesian Nash equilibrium.

5.5.3 Results

The pedestrian has consistent preference such that a potential collision for the pedestrian yields the worst payoff and she receives the highest payoff when the AVs yield for her in the interactions with outsiders[35] of AV protection priority AV (AVI) and insider's protection priority AV (AVII). So, when the two types of AV choose *Go*, the pedestrian's expected payoff to *Stop* is independent of p, the percentage of AVI on the road. Therefore, there is only one threshold[36], $\bar{p}$.

$$\bar{p} = \frac{e-a}{c-a} \quad - \quad (3)$$

$\bar{P}$ is the function of the pedestrian's payoffs in the situations for the potential collision (a), when the AVs yield for her (c), and when she yields for the AVs (e). It is the threshold probability of type I AV and it satisfies $0 \leq \bar{p} \leq 1$. Because the pedestrian's preference with both AVI and AVII is $c > e > a$[37] and the cost of potential collision for the pedestrian is non-positive, $a < 0$. Therefore, the pedestrian's preference is not only necessary, but also sufficient condition to estimate $\bar{p}$.

The general condition of a static Bayesian Nash equilibrium shows that if the proportion of outsider's protection priority AV (AVI) is less than $\bar{p}$ on the road, the pedestrian's best action is always to stop and yield for both types of AV, i.e., there is AV supremacy. In other words, if $p < \bar{p}$, one Bayesian Nash equilibrium exists: (Pedestrian, AVI, AVII) = (*Stop*, (*Go*, *Go*)), where the first component is the action of pedestrian and the other component is the pair of actions of the two different types of AV, *AVI* and *AVII*, respectively. In a Bayesian Nash equilibrium of (*Stop*, (*Go*, *Go*)), given that the actions of AVI and AVII are (*Go*, *Go*), the pedestrian's action *Stop* is optimal and given that the pedestrian chooses *Stop*, *Go* is optimal for both the AVI and the AVII. It means that when both the AVI and the AVII go, the pedestrian

[35] There are other members of outsider of vehicle such as cyclists, motorcyclists, and other human drivers. In this analysis, I only consider pedestrians for simplicity.

[36] It is named as the *ACE* threshold.

[37] The pedestrian condition (1), Section 5.2.2.

stops. It is an inefficient outcome in mixed transportations. However, if the proportion of AVI is more than or equal to $\bar{p}$ on the road, $p \geq \bar{p}$, there are two Bayesian Nash equilibria; (Pedestrian, AVI, AVII) = (*Stop*, (*Go*, *Go*)) and (*Go*, (*Stop*, *Go*)). For a Bayesian Nash equilibrium of (*Go*, (*Stop*, *Go*)), given that the actions of the AVI and the AVII are (*Stop*, *Go*), the pedestrian's action *Go* is optimal and given that the pedestrian chooses *Go*, *Stop* is optimal for the AVI but *Go* is optimal for the AVII. That is, the pedestrian goes, the AVI stops but the AVII goes. It leads the *Pareto* efficient outcome in mixed traffics such that the AVI yields for the pedestrian and the pedestrian goes while the insider's protection priority AV (AVII) also goes[38]. These two equilibria are compatible with a steady state; both outcomes are stable social norms[39]. In practice, the focal Bayesian Nash equilibrium depends on local social norms.

It is interesting to notice that in all equilibria, AVII always goes. That is, *regardless of whether the pedestrian and the AVI stops or goes*, AVII goes in the situation where it needs to protect the insiders of the vehicle. This highlights a current emphasis of designing and programming passenger safety AV for the car manufacturers and resolves a potential consumer's concern about the safety of an AV. In the meantime, introducing the AVI not only lessens the public outrage about only passengers' safety AV but also resolves the potential AV supremacy or pedestrian supremacy.

As the cost of potential collision with pedestrian, a rise, $\bar{p}$ rises. It says that society needs more types of AV that prioritize to protect the outsiders of AV (AVI), which makes sense.

As the pedestrian's payoff rises when AVs yield for her, the higher c, the lower $\bar{p}$ is such that the less type I AV is needed. As pedestrians as outsiders of AV are less happy with the interaction with AVs, the lower c, the society needs more type of AV that yields for pedestrians, the higher $\bar{p}$. That is, as pedestrians are less satisfied with AV's yielding behavior, society needs more types of AV that yields for pedestrians. This result makes sense, because as pedestrians are happier from AV's yield behavior, the mixed traffic condition is stable enough and the society does not need more AVI.

[38] People may argue about AVII's dominant strategy of *Go*. That is, the insiders protection priority AV goes even when pedestrian goes. This analysis considers the ethical dilemma situation, where facing unavoidable situations. Under normal situations, AVII will follow the traffic rules and stop for pedestrians at crosswalks and at traffic lights. From the other aspect, this result supports the current development of passenger safety AV such that it goes to protect passenger in unavoidable situations. So, the potential consumers of AV feel comfortable purchasing it.

[39] In some games with many Nash equilibria, some of these equilibria seem more likely to attract the players' attentions than others. It implies that there is preferable equilibrium, and they seem more likely to be *focal*.

The higher e means that the pedestrian's yield payoff rises. It indicates that there are already many AVII on the road, so the pedestrian prefers to yield for AVII. Otherwise, the pedestrian would collide with the AVII and cause the worst outcome for her. It implies that as the pedestrian's preference to yield rises, the AVII's dominant strategy '*Go*' leads to AV supremacy. Therefore, society needs more types of AV that prioritizes the protection of outsiders of the AV (AVI) to prevent AV supremacy.

As the pedestrian's yield payoff rises (the higher e), the more type I AV is needed (the higher $\bar{p}$) to prevent AV supremacy. This situation also can be regarded as such that the pedestrian's preference to yield rises, because there are more type II AVs (AVII) that prioritize the safety of the AV passengers. Therefore, she prefers to stop and yield for the AV. Otherwise, she would collide with the AV and cause the worst outcome for her. It implies that as the pedestrian yields more for AVI and AVII, AVII's dominant strategy '*Go*' leads to 'AV supremacy'. Therefore, to prevent this inefficient outcome, society needs a type of AV that prioritizes the protection of outsiders of the AV (AVI). This result also supports the human-centric society under the condition of the higher c and the lower e.

The existence of two different types of AV reconciles moral values and personal self-interest of human users and traffic participants by embedding different moral preferences to each of the two AV types. A static Bayesian game model to address heterogeneous moral preferences in an AV decision dilemma shows that as the outsiders protection priority AV (AVI) is introduced as a different type of AV than the insiders protection priority AV (AVII or DAV), and if certain portion of the AVI is available on the road, it leads to an efficient outcome where the pedestrian and the AVI yield each other while the AVII keep going. This game theoretic approach in AV dilemma also shows to prevent human traffic participant's moral hazard behavior and improve transportation efficiency in mixed traffic.

III. The Role of Government for Ethical AI
The Regulations of Moral Algorithms in AV systems:
From no single but multiple moral values in AV dilemma

6. The current AI ethics guidelines and policies

As AI's adoption grows more widespread and companies see increasing returns on their AI investments, the technology's risks become more apparent. One of the most transformational shifts has been with transportation and the transition to AVs. The evolution of AV technology raises a number of important legal and regulatory issues (OECD, 2019). Ensuring safety and liability are primary concerns. Floridi et al. (2018) indicate that there is risk that AI may lead to unplanned and unwelcome changes in human behaviors arising in response to accommodating the routines that make people's lives easier. As the role of government is critical for addressing the ethical concerns and legal challenges, it is imperative that more research must be carried out on the role of the government in shaping the future of AI. Coeckelbergh (2019) states that most of ethical principles and concepts are usually formulated and addressed in an abstract and general way. The question of methods and implementations in AI systems remains unclear what exactly we should do.

Many people are unaware how AI systems are being used to make decisions that impact their lives in the United States (EPIC, 2023). Accordingly, lawmakers are developing the meaningful AI relevant legislation at both national and international levels (EPIC, 2023; OECD, 2019, 2023; *The Universal Guidelines for Artificial Intelligence*). Electronic Privacy Information Center (EPIC) urges government to use the universal guidelines for AI[40] (*The Universal Guidelines for Artificial Intelligence*). The OECD AI Principles[41] (OECD, 2023) establish international standards for AI use such as human-centred values and fairness, and transparency and explainability. They provide frameworks to guide their policymaking towards equitable solutions. All regulations are required to recognize the harm posed by AI systems. In particular, they require transparency and explainability for both commercial and government uses of AI[42] as well as at national and international levels (OECD, 2019, 2023). Transparency and explainability indicate that all individuals have the right to know and understand the basis

[40] It is endorsed by more than 40 countries, over 250 experts and 60 organizations in October 2018.

[41] The OECD AI Principles were adopted in 2019 and endorsed by 42 countries, including the United States, several European Countries, and the G20 nations.

[42] Electronic Privacy Information Center (EPIC) introduces specific issues related to AI and Human Rights, such as AI in the criminal justice system, commercial AI use, and government AI use (EPIC, 2023).

of AI-based outcomes that concern them. This includes access to the factors, the logic, and techniques that produced the outcome (EPIC, 2023; *The Universal Guidelines for Artificial Intelligence*). Organisation for Economic Cooperation and Development (OECD) encourages public and private investments in AI research and development to spur innovation in trustworthy AI that focus on AI-related social, legal and ethical implications and policy issues (OECD, 2023). It is well summarized how several different stakeholder groups are already actively engaged in discussions on how to steer AI development and deployment to serve all of society (OECD, 2019, p. 123).

The United Kingdom government pursue a new pro-innovation approach to regulating AI in a way to balance between unleashing the full potential of new AI technologies and safeguarding human's fundamental values (UK, 2022). It is a context- and risk-based approach which is flexible and preferable to a single framework. The United Kingdom government claims that a centralized approach with a fixed list of risks may stifle innovation. The United Kingdom government does not think that the EU's approach in the product safety regulation of the single market captures the full application of AI and its regulatory implications. A context-based approach proposes to regulate AI based on its use and the impact it has on individuals, groups, and businesses within a particular context. A risk-based approach focuses on addressing issues where there is clear evidence of real risks associated with AI. The United Kingdom government want this alternative approach to be adaptable to AI's vast range of uses across different industries, and support regulators in addressing new challenges in a way that encourages innovation and avoid unnecessary barriers. While such an approach would offer maximum flexibility, it raises the risk that businesses and the public would not have a consistent view of what is and is not the subject of regulation. So, the United Kingdom government intend to set out the core characteristics of AI to inform the scope of the AI regulatory framework but allow regulators to set out and evolve more detailed definitions of AI according to their specific domains or sectors.

The policies are also targeted at enabling the cooperative work of humans and AI (OECD, 2019, 2023). As AI technology diffuses, the potential impacts of its prediction, recommendations on people's lives increases. The technical, business and policy communities are actively exploring how best to make AI human-centred and trustworthy, maximize benefits, minimize risks and promote social acceptance. AI actors should provide meaningful information, appropriate to the context, and consistent with the state of art to make stakeholders aware of their interactions with AI systems and understand the outcome by an AI system.

Above all, much work in AI ethics in recent years has focused on developing high-level principles. But these principles and guidelines say nothing about what to do when principles come into conflict with one another. For example, principles do not tell how to balance the potential of AI to save lives (the principle of beneficence) against other important values such as privacy or fairness (Tzachor et al., 2020).

> Many of the real-world moral codes one might wish to implement in a computational system do not prioritize all the rules and are therefore subject to conflicts between them. Any rule-based AMA will require a software architecture that can manage situations in which rules conflict (Wallach & Allen, 2009, p. 93).

Schmidt (2021) claims that the global technology competition is ultimately a competition of values, and concerns about the impact of AI development on the conflicting values between democratic and authoritarian systems. According to Schmidt (2021), the competition of AI developments and its employments may underpin the current contest of values between democracy and authoritarianism on individual liberty and human rights.

Etzioni and Etzioni (2017) claim that moral values are implemented through legal enforcement and personal choices. Humans can provide moral guidance to autonomous smart machines through legal enforcement, a very familiar process of law-and regulation-making for collective decision. As to those decisions left open-ended by lawmakers, the response is left to each individual. In the case of driverless cars, driverless cars have to obey the law like all other cars. The ethical decisions that are not prescribed by law are left to be made by individuals or their cars. A significant part of the ethical challenges posed by AI-equipped machines can be addressed by the kind of ethical choices made by humans. That is, for individuals to instruct the car they own or use to follow their value preferences. The challenge is to find ways for owners or users of driverless cars to guide them in these matters, but not for ethically mandated choices by some collective or third party.

Renieris et al. (2022) claim that companies leading the way on a responsible AI are not driven primarily by risks, regulations, or other operational concerns. Rather, leaders take a strategic view of responsible AI, emphasizing their organizations' external stakeholders, broader long-term goals and values, and social responsibility. In short, responsible AI actually has less to do with AI than with organizational culture, priorities, and practices. It is about how the organization views itself in relation to internal and external stakeholders, including society as a whole.

This chapter explores public policy considerations to ensure that AI systems are trustworthy and human-centred. The need to progress towards more robust, safe, secure and transparent AI systems with clear accountability mechanisms for their outcomes is underlined.

7. The practical and effective solutions of moral algorithms in AV systems

> A more productive way to think about the problem … such that the overall outcome would still be acceptable, if you couldn't tell him what specifically he was doing wrong? … We should not try to invent a "super" version of what our own civilization considers to be ethics (Bostrom & Yudkowsky, 2018, p. 17).

The development of AVs raises ethical concerns about how to program an AV to make moral decisions in dilemma situations where there is a risk of harm to passengers, pedestrians, or other drivers. To resolve these ethical dilemmas, it is recommended to develop clear ethical guidelines and frameworks that prioritize safety, fairness, and respect for human life. It demands engineers and designers of AV to incorporate ethical considerations into the design process of AVs, such as designing algorithms that prioritize the safety of all parties involved. It is also recommended to develop legal and regulatory frameworks that govern the use of AVs. It is a reminder that there is an ethical dilemma in the context of protecting different stakeholders, such as pedestrians and passengers of AV, from possible harm in AV settings. The dilemma arises because it is not always possible to keep every traffic participant equally safe. In the ethics of AV, it is a choice of who to protect: an AV has to choose which stakeholder's safety is to be prioritized in dilemma situations. In the literature on the ethics of an AV, the trolly problem and its variant examples[43] has been discussed to demonstrate that different stakeholders reveal different moral preferences for the choice of AV in protecting them from harm. Bonnefon (2016) regards it as the *tragedy of the commons*[44] and comments that enforcing regulations may provide a solution to these problems. However, regulators will face some difficulties. It is shown that most people appear to be reluctant to accept government regulations imposing utilitarian algorithms[45] in AVs (Bonnefon et al., 2016). Regulation for

[43] Under what kinds of circumstances a life-and-death decision would or would not be permissible by an AV, and what self-driving cars should be programmed to do in various crash dilemma scenarios.

[44] People acting in their self-interest behave contrary to the actions that everyone knows are necessary for the common good. In the context of heterogeneous moral preference of AV, a large majority of participants would like to see "utilitarian" AV on the road, but most people also indicated that they would refuse to purchase such an AV. Rather they express a strong preference for buying the "self-protective" AV. (Bonnefon, 2016).

[45] Utilitarian AVs sacrifice their passengers for the greater good and minimize the number of total casualties.

utilitarian algorithms may postpone the adoption of AVs. Bonnefon et al. (2016) state that "if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs" and suggest that car makers and regulators should considering solutions to these obstacles. Zhu et al. (2022) claim that it seems unlikely that AVs will be programmed with a one-size-fits-all algorithm due to different moral preferences in AV dilemma.

7.1 Floating conclusions approach in AV dilemma[46]

Floating conclusion approach as skeptical reason over conflicting propositions explicitly describes a dilemma in a way that two source of information infringes each other (Bonnefon, 2004; Horty, 2002). Acceptable floating conclusions[47] from conflicting principles or values enable to embed conflicting ones independently in the AI system. These support the existence of two different types of autonomous vehicle (AV) reflecting different ethical principles and moral values to address ethical issues in machine ethics.

In addition to carrying passengers to the desired destination, safe, fuel-efficient and travel time efficient as same optimization goals (Nyholm & Smids, 2020), breaking the straight line can be the other *acceptable* common goal for both AVI and AVII. Because it is the safest way rather than swerve in dilemma situations (Davnall, 2020). According to Davnall (2020), swerving is less attractive than the emergency stop to brake as hard as possible in a straight line, because of its unpredictability and the fact that it results in higher speed impacts. Swerving generates a huge number of risks in the case of AV. This analysis can be structured in the extensive form of floating conclusion approach as follows in Figure 3, where

$$\begin{cases} T: \text{The existence of two types of AV} \\ A: \text{Type I AV} \\ B: \text{Type II AV} \\ c: \text{Outsiders protection priority} \\ -c: \text{Insiders protection priority} \\ D: \text{Breaking the straight line} \end{cases}$$

T is the observation such that two types of AV – type I AV (A) and type II AV (B) exist whose protection priority conflict with each other; outsiders protection priority (C) versus insiders protection priority ($-C$). Yet, breaking the straight line is justified as an acceptable floating conclusion (D) from conflicting protection priority in two distinct types of AV.

[46] I also thank Dr. Rune Nyrup for helpful discussions about this idea.

[47] See Section 2.

The interpretation of the arrows[48] is in a way that $A \rightarrow (C \cap D)$ and $B \rightarrow (-C \cap D)$. Therefore, $A \subseteq (C \cap D)$ and $B \subseteq (-C \cap D)$, where D is an acceptable floating conclusion without loss of generality.

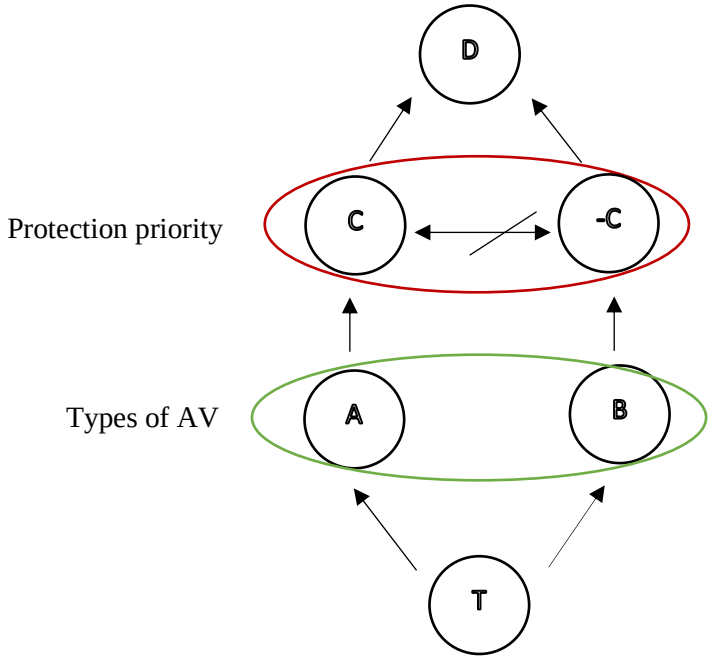


Figure 3
Breaking the straight line as an acceptable floating conclusion in AV dilemma

Etzioni and Etzioni (2017) suggest that enhanced moral guidance to smart machines should draw on a new AI program that will read the owner's moral preferences and then instruct these machines to heed them– *ethics bot*.

> Many of the ethical decisions that smart machines are said to have to make, need not and should not be made by them. Because they are entrenched in law. These choices are made for the machines by society, using legislatures and courts. Many of the remaining ethical decisions can be made by ethics bots, which align the cars' conduct with the moral preferences of the owners. Neither the law nor ethics bots can cover extreme outlier situations (p. 415).

A floating conclusion approach in AV dilemma helps to implement (conflicting) moral values through personal moral preferences and choices. This approach covers the ethical decisions that are not prescribed by law in a way that enables *ethics bots* to embed two different moral principle, extract specific ethical preference from a user, and subsequently applies these preferences to the operations of the user's AV in dilemma cases. It provides a way for owners or users of driverless cars to guide their AVs in these matters, but not for ethically mandated

[48] There is no distinction between logical and default implication of arrows. See Section 2.

choices by some collective or third party. The government focuses on collective ethical choices by pursuing acceptable common conclusions from conflicting propositions in line through legal means.

7.2 A Static Bayesian Game in AV dilemma

A static Bayesian game model[49] shows that not only the existence of outsiders protection priority AV (AVI) as another different type of AV than insiders protection priority AV (AVII), but also the maintenance of certain portion of AVI (*P*) led to the *Pareto* efficient outcome in mixed traffics.

The general condition of a static Bayesian Nash equilibrium is summarized below in Table 8.

	P	
	$0 \leq p < \frac{e-a}{c-a}$	$\frac{e-a}{c-a} \leq p \leq 1$
Bayesian Nash equilibrium	(*Stop*, (*Go*, *Go*))	(*Stop*, (*Go*, *Go*)) (*Go*, (*Stop*, *Go*))

Table 8 The general conditions of static Bayesian Nash equilibrium

where

$$\begin{cases} a\text{: pedestrian's payoff in the potential collision} \\ c\text{: pedestrian's payoff when AVs yield for her} \\ e\text{: pedestrian's payoff when she yields for AV} \end{cases}$$

It is reminded that a is the pedestrian's payoffs in the potential collision with AVs, c is the pedestrian's payoffs when the AVs yield for her, and e is the pedestrian's payoffs when she yields for the AVs. The pedestrian's preference with both AVI and AVII is $c > e > a$ and $a < 0$.

Policymaker's aim is to find the optimal proportion of AVI ($\bar{p}$) for the safer environment in mixed traffic. Policymakers do not want to have all AVs are either AVI ($\bar{p} = 1$) or AVII ($\bar{p} = 0$). Policymakers may want p as much as close to 1 as possible but less than 1 for the improvement of overall safety and transportation efficiency.

$$\bar{p} = \frac{e-a}{c-a}$$

[49] See Section 5.5.3.

If policymakers focus on policy tools that lower $\bar{p}$, the social planner does not need to purchase many AVI to achieve the *Pareto* efficient outcome.

The higher c is, the less AVI is needed: the pedestrians are happier from AV's yielding behavior because there are "enough" numbers of AVI to prioritize the pedestrians' safety as outsiders of the AV in mixed traffic conditions. Therefore, society does not need more AVI. If not, it causes pedestrian supremacy. The lower c is, the more AVI is needed: as the pedestrian as outsiders of the AVs are less satisfied with AV's yielding behavior, it implies that these AVs are the AVII that prioritizes the safety of AV. So, society needs more types of AV that yields for the pedestrians, AVI. More AVI is needed to prevent the self-protective AV (AVII) supremacy.

The lower e means that the pedestrian's yield payoff falls. It indicates that there are already many AVI on the road, so the pedestrian prefers not to yield for AVI. A decrease in the pedestrian's preference to yield implies that '*Stop'* becomes AVI's dominant strategy, and it leads to pedestrian supremacy. Therefore, society needs less type of AV that prioritizes the protection of outsiders of the AV (AVI) to prevent pedestrian supremacy.

Society needs less AVI with a lower cost of potential collision with pedestrians.

If the social planner needs to buy AVI for the overall safety environment in mixed traffics, how could it be used for? Lin (2013) raises the ethical issue of a publicly owned AV. The owner of AV may reasonably expect that its property "owes allegiance" to the owner and should value his or her life more than unknown pedestrians and drivers. But a publicly owned AV might not have that obligation, and this can change moral calculations. The duties of AV may vary, so even among public AV, the assigned roles and responsibilities are different between them. Some AVs may be obligated to sacrifice themselves and their occupants in certain conditions, while others are not.

Sparrow and Howard (2017) raise concerns about the ethical aspect of driverless vehicles that are appealing to consumers, as well as the policies that are likely to emerge onto the market. Sparrow and Howard (2017) believe that the safest way to bring a driverless vehicle to market would be to design and manufacture a vehicle that possesses SAE level 4 (or 5) autonomy[50]. At this high automation level, the autonomous vehicle monitors the driving environment and

[50] The Society of Automotive Engineers (SAE) defines 6 levels of driving automation range from 0 (fully manual) to 5 (fully autonomous). SAE Level 4 and 5 describe the situations such that "You (the human) are not driving when these automated driving features are engaged – even if you are seated in "the driver's seat". These automated driving features will not require you to take over driving." (Source: www.sae.org/blog/sae-j3016-update)

performs all driving tasks under specific (level 4) and all (level 5) conditions. The OECD report mentions about the availability of the technology and the role of the driver (OECD, 2019). As AV technology is advanced, some firms such as Tesla and Waymo believe that it will be possible to eliminate the need for a human driver. Tesla sells with level 3 autonomy[51]. Waymo had plans to launch a fully autonomous taxi service with no driver in Arizona in the United States. Other systems developers believe that the best use of AV systems for the near term will be supporting the driver rather than replacing drivers. Toyota is emphasizing development of vehicles that are incapable of causing a crash.

The role and use of AVI in society is discussed as follows.

7.2.1 The role and use of AVI

The perception and recognition of the AVII prevents road users' bad behavior to exploit AV's safe and predictable technology. Yet, it may cause AV supremacy. An AVI works as the "Pareto efficient maker" in a way that the AVI stops for outsiders of AV, including pedestrians in unavoidable dilemma situations. However, the existence of AVI is insufficient to improve the Pareto efficient outcome in mixed traffic. It requires to maintain a certain proportion of AVI ($\bar{p}$) in mixed transportation.

The social planner purchases a certain number of AVI to maintain the optimal level of $\bar{p}$ and use it for the public services such as cleaning the street[52]. Today's new technologies, innovations, the abundance of data and its digitized information are creating advantageous entrepreneurship environments such as internet-based platforms. The internet-based platforms shift business environments and enable the growth of the gig and sharing economy. The sharing economy not only presents a new economic paradigm, but also provides the local government an insight to explore new revenue sources. The local authorities own the AVI for sharing cars or public transportation to tourists who are unfamiliar to their local traffic customs and rules. For example, the position of driver seat in vehicles is the opposite in the United Kingdom compared to in Italy. Italian tourists may ride in the AVI to travel the United Kingdom while adjusting themselves into new traffic environments.

An AVI can be used as driverless "designated driver" who drives drunk customers when they have had too many drinks to drive safely. The idea of Drive Home service is already

[51] SAE Level 3 describe the situations such that "You (the human) are not driving when these automated driving features are engaged – even if you are seated in "the driver's seat"." However, "when the feature requests, you must drive." (Source: www.sae.org/blog/sae-j3016-update)

[52] In this case, an AVI could be "driverless" cleaning vehicle.

available and people use it (Fowler, 2015). Yet, Fowler (2015) pointed out whether an extremely drunk person would have had used such designated driver service easily. According to the National Highway Traffic Safety Administration, "every day, one person every 45 minutes dies in drunk-driving crashes in the United States. In 2020, 11, 654 people died in alcohol-impaired driving traffic deaths – a 14% increase from 2019." (NHTSA, 2020). The use of AVI as the chauffeur service to drunk people will prevent drunk driving and improve road safety.

An AVI also can be used to balance the safety of the drivers and other road users with the rights of elderly drivers to maintain their independence and mobility. The growing ageing population raises a question about elderly people's driving. It presents additional requirements to promote safe driving among older people while recognizing the potential risks associated with aging and driving. Different countries have different policies and regulations regarding elderly drivers. These policies may vary depending on the age of the driver and other factors such as health status and driving ability. For example, in Japan, drivers over the age of 75 are required to undergo cognitive and physical tests every three years. In the United Kingdom, drivers over the age of seventy must renew their driver's license every three years and are required to declare any medical conditions that may affect their driving abilities. If there are concerns about their ability to drive safely, they may be also required to take a medical or driving test to renew their licenses (ChatGPT, March 27, 2023).

Now, if a manufacturer offers different versions of its moral algorithm in AV such as AVI and AVII, and a buyer knowingly chose one of them, is the buyer to blame for the harmful consequences of the algorithms' decisions? Such liability consideration needs to be discussed.

7.2.2 The randomization of AVI allocation

Bonnefon et al. (2016) state that "if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs." A similar argument arises that if both AVI and AVII were available on the market, few potential owners and users of AV would be willing to ride in AVI. Because AVI is prioritized to protect the outsiders of AV.

The randomized allocation of the AVI and AVII to the potential consumers of AV is considered to resolve liability issue. Policymakers allow manufacturers to produce AVI and AVII while maintaining $\bar{p}$. Then, no explicit choice option of AVI and AVII is available to the potential consumers. It implies that when the potential customers decide to purchase an AV,

the type of AV is randomly assigned and sold to them. The car manufacturers can explain their operating ethical principles to them if they ask. This approach is relevant to responsibility and liability issues in AV dilemma situations as follows. If a manufacturer offers different versions of its moral algorithm with AVI and AVII, and a buyer knowingly chooses one of them, there is question whether the buyer is to blame for the harmful consequences of algorithms' decisions or not. But with the randomized allocation of AV, the owners and users of AV become less liable. This incentivizes them to accept the randomized AV.

The randomization of different type of AV allocation also relates to the effect of market forces and consumer preferences. Bonnefon et al. (2019) provide an example such that two competing companies market AVs that both eliminate 80% of fatalities. But one company's AVs split the remaining fatalities equally between insiders (passengers) and outsiders (pedestrians) of AV, whereas the other company's AVs split the remaining fatalities nine-to-one in favor of their passengers. Consumers would flock to the second company, and pedestrian risks would gradually inflate to unacceptably unfair levels and lead to AV supremacy.

7.3 Policy Implications

Sparrow and Howard (2017) claim that governments will have strong ethical, public policy, and economic reasons to incentivize the uptake of driverless vehicles once their introduction reduces the road toll. However, frameworks that incorporate ethical considerations into agents which are designed to influence human behaviors remain open for programmers and designers of AI.

Hagendorff (2020) review 22 of the major AI ethics guidelines and examine to what extent the ethical principles and values are implemented in the practice of research, development and application of AI systems. Hagendorff (2020) claims that very little to nothing has been written about the tangible implementation of ethical goals and values in the field of AI. Given the relative lack of tangible impact of the normative objectives set out in the guidelines, the question is how the ethical guidelines could be improved to make them more effective. It is challenging to deduce concrete technological implementations from the very abstract ethical values and principles. For example, what does it mean to implement justice or transparency in AI systems? What does a "human-centered" AI look like? Hagendorff (2020) states that ethics lacks a reinforcement mechanism. Therefore, it is necessary to build tangible bridges between abstract values and technical implementations in AI systems.

Greene (2014) shows that our moral brain is a dual mode with automatic settings as efficient but inflexible and a manual setting as a flexible but inefficient. Therefore, if we want to develop

ethical AV as human-like or human-centred, it is reasonable to adopt conflicting moral principles such as deontic and utilitarian rules as a dual mode in AV systems as well.

The finding from a floating conclusion approach in dilemma in Section 2 indicates that multiple conflicting principles or goals can be implemented with an acceptable common conclusion from those conflicting propositions. It supports to program and develop two different types of AV that feature different philosophical and ethical specifications, and shows how to handle ethical dilemma in AV in Section 7.1. A Bayesian game model to address interactions between human subjects and AV provides a theoretical frame to develop feasible, practical and effective regulations for deploying intelligent systems and robots in the context of transportation in Section 5. These also provide policymakers with a framework of mechanism design to influence the human road users' payoffs and the use of AVI for a smooth human-AI collaboration in the AI age in Section 7.2.

> A successful moral AI system does not necessarily have to dictate one true answer in such cases (Conitzer et al., 2018, p. 4382).

No single but multiple ethical principles allow policymakers to develop the mechanism design to resolve the ethical dilemma in AV systems. Finding the single ethical principle or moral value to address and resolve ethical issues in AI and AV is difficult and challenging as discussed before. Rather than choosing one principle among several conflicting ethical principles, let multiple conflicting ethical principles be there and allow the stakeholders to choose the relevant multiple principles with *acceptable common conclusions* from those conflicting ones. It incentivizes manufacturers to advertise their operating ethical principles in AI and aligns with the United Kingdom government's pro-innovation approach[53] to regulating AI (UK, 2022).

8. Further Research

When an AI system makes a decision that causes unintended harm, who is to be blamed – the programmer who developed the underlying algorithm or the manufacturer who produced AV or the customer who purchased and used the system? Given the huge penalties that may be associated with liability lawsuits, this issue is probably one of the most pressing ones in need of legal clarification.

[53] See Section 6.

As the AV technology is advanced to learn human road users' traffic ethical behavior and update it as an explicit player, repeated game, particularly a dynamic Bayesian game model can be considered to analyze the strategic interaction between pedestrians and two types of AV.

8.1 Responsibility and Liability in AV Ethics

Grinbaum et al. (2017) regard the liability problem as authority sharing problem, because actual decision making will be shared between the robot and the operator. Grinbaum et al. (2017) recommend researchers to decide whether a human or a robot holds the decision-making power at a given time. If all the decisions of a human can possibly be delegated to a robot, for example, when a quick reaction is necessary or in the absence of communication with the operator, researchers must address the issue of reliability of the knowledge and algorithms underlying a robot's decisions and its limits[54].

Crawford et al. (2019) mention about ethical responsibility in a way that contemporary AI systems perform a diverse range of activities and poses new challenges to traditional ethical frameworks due to the implicit and explicit assumptions made by these systems. When these systems are deployed in human contexts, potentially unpredictable interactions and outcomes occur.

The case of airplane crash liability can be considered in the case of AV liability, because there is autopilot system on airplane. Autopilot systems maintain altitude and are incredibly consistent. According to *Airplane Crash Liability Overview* (2016), most pilots like to hand-fly until an altitude of around 10,000 feet. That is, pilots must wait until the plane has reached an altitude of at least 400 – 10,000 feet before turning on autopilot. For the entirety of the takeoff, the pilot is hand-flying and is in control. This goes for taxing around the airport as well. After going over 28,000 feet, it is mandatory for pilots to have autopilot engaged. If the weather condition is bad such as heavy fog, pilots will often turn the autopilot system on much sooner after taking off. Because when it is difficult to see, the autopilot system does not require eyesight like humans do and it can be extremely helpful. This makes it a vital part of any successful flight or landing in the most severe weather conditions. Current regulations state that autopilot must be disengaged upon reaching the decision altitude (DA). Most commonly this altitude is 200 feet above the touchdown area, and this is applicable during most landings.

[54] In the context of endowing moral behavior into a robot, there are programming limits. So researchers must evaluate whether general rules are applicable, if the notion of the right action is relevant to the moral framework used in computation, and how moral values are ranked in controversial decision making. (Grinbaum et al., 2017).

The majority of pilots let autopilot handle the controls until the landing area and runway is clearly in sight and visible.

Pilot error and mechanical failure are the leading causes of airplane crashes. When pilot error, maintenance deficiencies, or other lapses cause a crash, a lawsuit can claim that the airline's negligence was responsible[55]. If manufacturers of airplanes and component parts cause a crash, manufacturer's strict product liability is applied (*Airplane Crash Liability Overview*, 2016). Kaplan and Haenlein (2020) propose that self-driving cars might be required to have black boxes, like flight recorders, which can be used to provide objective information in case of an accident.

The randomized allocation of two different types of AV to the potential customers who purchased AV is suggested to consider liability issue in AV dilemma situations[56]. However, in case of an accident caused by AV in a clear and obvious situation[57], it is still unclear who is responsible – the car manufacturer who produced AV, the software programmer who developed the underlying algorithm, or the user of the vehicle who was not driving – for damage, injury, or loss.

The fair distribution of liability among outsiders[58] of AV is also needed to further discuss. Although an AVI is prioritized to protect pedestrians as a member of outsider of AV, the AVI's protection priority to the pedestrian needs not to be a perfect no harm guarantee for her. The risk of potential collision with AVI still exists for pedestrian[59].

8.2 Dynamic Bayesian Game Model in AV dilemma

An AV is assumed to be an implicit ethical player who is unable to learn and update encoded ethics like pedestrian[60]. This characterizes a static game. If AV has a learning system

[55] It is airlines' liability.

[56] See Section 7.2.2.

[57] It describes non-dilemma situation that may be easier to deal with than dilemma situation. However, even in a clear and obvious situation, if an AV makes a decision that causes an accident, determining responsibility can be difficult and may require a redefinition of liability laws. According to Kaplan & Haenlein (2020), given the huge penalties that may be associated with liability lawsuits, the liability issue in AV is probably one of the most pressing ones in need of legal clarification.

[58] Pedestrians, cyclists and motorcyclists are assumed to be the members of outsider of AV in Section 4.

[59] It can be regarded as the shift of relative risk from one road user to another (Bonnefon et al., 2019) within outsiders of AV.

[60] See Section 5.2.1.

that enables it to learn human road users' traffic ethical behavior and update it as an explicit player as pedestrian, the dynamic Bayesian game model could be applied to analyze the strategic interaction between pedestrians and two types of AV. Whether two static Bayesian Nash equilibria would be held in repeated games setting will be an interest of research.

Bonnefon et al. (2019) mention about the dynamic concept of AV dilemma in statistical perspective. It is claimed that trolley dilemmas are merely the unrealistic discrete version of a very real dilemma that emerges at a statistical level. It means that the trolley dilemma and the statistical trolley dilemma are equivalent. Because once unrealistic and improbable cases of the trolley dilemma are aggregated over millions of cars driving billions of miles, these small statistical decisions add up to life and death consequences – and prompt the same questions as the trolley dilemma did. This statistical trolley dilemma needs to be solved.

Conclusions

The development of AI and its impact on our societies raise concerns about the way AI and automated decision-making systems cause intended or unintended outcomes. Some of ethical problems are unintended consequences of AI's unique technology. For every ethical principle, there appear to be moral trade-offs. Humans confront such dilemmas every day. If all relevant ethical principles and values are considered equally in the ethical issues of AI, AI needs to use human's moral criteria to rank different courses of action and choose the one that is preferable to ones. The key ethical values and principles may conflict with each other to embed them in AI systems. Yet, there is few research about how to resolve two or more conflicting ethical principles or moral values and explain the logic and ethical decision making in AI system meaningfully.

Floating conclusion as a logical tool and skeptical reason over conflicting propositions is proposed to explicitly describe dilemma situations. Acceptable floating conclusions is justified as a common conclusion from extensions associated with conflicting propositions. A floating conclusion approach provides a framework to conceptualize conflicting ethical principles that can be extended to the ethics of AI where ethical principles and moral values conflict with each other in the AI system. A floating conclusion approach in dilemma indicates that multiple conflicting principles or goals can be implemented with an acceptable common conclusion from those conflicting propositions.

According to Bonnefon (2016),

> Car manufacturers have generally remained silent on the matter (of AV ethical dilemma). That changed when an official at Mercedes-Benz indicated that in those situations where its future autonomous cars would have to choose between risks to their passengers and risks to pedestrians, the algorithm would prioritize passenger safety. But the company reversed course soon after, saying that this would not be its policy... Carmakers can either alienate the public by offering cars that behave in a way that is perceived as unethical, or alienate buyers by offering cars that behave in a way that scares them away. In the face of this, most car companies have found that their best course of action is to sidestep the question: ethical dilemmas on the road are exceedingly rare, the argument goes, and companies should focus on eliminating rather than solving them.

In the case of AV, an acceptable floating conclusion from extensions associated with conflicting ethical principles or moral values enables to embed conflicting ones independently and accept them as extensions to support a common conclusion in AV settings. It reconciles heterogeneous moral values and shows how to handle ethical dilemmas in AV. It considers multi-stakeholder's perspective to provide practical mechanism to effectively regulate AI in AV settings. A floating conclusion approach supports programming two different types of AV that feature different philosophical and ethical specifications. Car manufacturers can produce different types of ethical driverless cars with passenger protection priority and pedestrian protection priority in their AVs. They have an incentive to advertise their operating ethical principles and it enhances transparency.

A new type of two reasoners in AV decision – insiders and outsiders of AV is proposed to encode the conflicting moral values and integrate the interests of different parties involved in AV environment. It enables car manufacturers to develop two different types of AV, the outsiders– and insiders – protection priority AV, AVI and AVII respectively that covers broad human road users, such as pedestrians, cyclists, and motorcyclists in mixed traffic. It provides the appropriate framework of the actual impact on individuals and groups in a particular context. The existence of two different types of AV reconciles moral values and personal self-interest of human users and traffic participants by embedding different moral preferences to each of the two AV types.

A static Bayesian game model is designed to address strategic interactions between pedestrian and two different types of AV. It shows that as the outsiders protection priority AV (AVI) is introduced as a different type of AV than the insiders protection priority AV (AVII), and if certain portion of the AVI is available on the road, it leads to an efficient outcome where the pedestrian and the AVI yield each other while the AVII keep going. It is interesting to notice that regardless of whether the pedestrian and the AVI stops or goes, AVII goes in the situation

where it needs to protect the insiders of the vehicle. This highlights a current emphasis of designing and programming passenger safety AV for the manufacturers and resolves a potential consumer's concern about the safety of an AV. This game theoretic approach in AV dilemma also shows to prevent human traffic participant's moral hazard behavior and improve transportation efficiency in mixed traffic. It supports the United Kingdom government's pro-innovation approach[61] in a way that AI is appropriately transparent and explainable in AV settings. This also helps to improve understanding of AI ethical decision-making in dilemma situations.

The policy implication – rather than choosing one principle among several conflicting ethical principles, let multiple conflicting ethical principles be there and allow the stakeholders to choose the relevant multiple principles with acceptable common conclusions from those conflicting ones – is derived from the finding with logical approach in AV dilemma.

No single but multiple ethical principles reconcile heterogeneous moral values in ethical dilemma situations. It incentivizes car manufacturers to advertise their operating ethical principles in their AVs. A game-theoretic approach to address interactions between human subjects and AVs provides policymakers a frame to develop feasible, practical and effective mechanism designs for deploying intelligent systems and robots in the context of transportation. These approaches align with the United Kingdom government's pro-innovation approach to regulating AI in a way to encourage innovation and avoid placing unnecessary barriers.

[61] See Section 6.

References

(ECB), T. (2022). *Monetary policy decisions*. https://www.ecb.europa.eu/press/pr/date/2022/html/ecb.mp221027~df1d778b84.en.html

Airplane Crash Liability Overview. (2016). FindLaw. https://www.findlaw.com/consumer/travel-rules-and-rights/airplane-crash-liability-overview.html

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc Natl Acad Sci U S A*, *117*(5), 2332-2337. https://doi.org/10.1073/pnas.1911517117

Baker-Brunnbauer, J. (2021). Management perspective of ethics in artificial intelligence. *AI and Ethics*, *1*(2), 173-181. https://doi.org/10.1007/s43681-020-00022-3

BBC. (2020). *Uber's self-driving operator charged over fatal crash*. https://www.bbc.com/news/technology-54175359

Bonnefon, J.-F. (2004). Reinstatement, floating conclusions, and the credulity of Mental Model reasoning. *Cognitive Science*, *28*(4), 621-631. https://doi.org/https://doi.org/10.1016/j.cogsci.2004.03.002

Bonnefon, J.-F. (2016). *Whose Life Should Your Car Save?* https://www.tse-fr.eu/article/whose-life-should-your-car-save

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576. https://doi.org/doi:10.1126/science.aaf2654

Bonnefon, J., Shariff, A., & Rahwan, I. (2019). The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars [point of view]. *Proceedings of the IEEE*, *107*(3), 502-504. https://doi.org/10.1109/JPROC.2019.2897447

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, *20*(1), 41-58. https://doi.org/10.1007/s10676-018-9444-x

Bosch, K. v. d., & Bronkhorst, A. W. (2018). Human-AI cooperation to benefit military decision making.

Bostrom, N., & Yudkowsky, E. (2018). The Ethics of Artificial Intelligence. *Artificial Intelligence Safety and Security*.

Bower, J. L. (1965). The Role of Conflict in Economic Decision-Making Groups: Some Empirical Results. *Quarterly Journal of Economics*, *79*, 263-277.

Bowles, S. a. H., Simon D. (2021). *Microeconomics: Competition, Conflict, and Coordination*. Oxford University Press.

Braithwaite, R. B. (1955). *Theory of games as a tool for the moral philosopher. An inaugural lecture delivered in Cambridge on 2 December 1954*. University Press.

Coeckelbergh, M. (2019). *AI ethics*. The MIT Press.

Commission, E. (2021). *REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL: LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. EUROPEAN COMMISSION. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. F. (2018). Moral Decision Making Frameworks for Artificial Intelligence. ISAIM,

Crawford, K., Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas,, Amba Kak, V. M., Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida, & Richardson, J. S., Sarah Myers West, and Meredith Whittaker. (2019). *AI Now 2019 Report*. N. Y. A. N. Institute. https://ainowinstitute.org/AI_Now_2019_Report.pdf

Davnall, R. (2020). Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. *Science and Engineering Ethics*, *26*(1), 431-449. https://doi.org/10.1007/s11948-019-00102-6

de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human Cooperation When Acting Through Autonomous Machines. *Proceedings of the National Academy of Sciences*, *116*(9), 3482-3487. https://doi.org/10.1073/pnas.1817656116

Di, X., Chen, X., & Talley, E. (2020). Liability design for autonomous vehicles and human-driven vehicles: A hierarchical game-theoretic approach. *Transportation Research Part C: Emerging Technologies*, *118*, 102710. https://doi.org/https://doi.org/10.1016/j.trc.2020.102710

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1-3. https://doi.org/10.1007/s10676-018-9450-z

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2019.01.021

EPIC. (2023). *AI & Human Rights*. Electronic Privacy Information Center (EPIC). https://epic.org/issues/ai/

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, *21*(4), 403-418. https://doi.org/10.1007/s10892-017-9252-2

European, C., Directorate-General for, R., & Innovation. (2020). *Ethics of connected and automated vehicles : recommendations on road safety, privacy, fairness, explainability and responsibility*. Publications Office. https://doi.org/doi/10.2777/966923

Fed, T. (2022). https://www.federalreserve.gov/aboutthefed.htm

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Fowler, G. A. (2015). *Tap Your Phone, and a Designated Driver Takes You (and Your Car) Home*. The Wall Street Journal. https://www.wsj.com/articles/tap-your-phone-and-a-designated-driver-takes-you-and-your-car-home-1435174499

Gawronski, B., & Beer, J. S. (2017). What makes moral dilemma judgments "utilitarian" or "deontological"? *Soc Neurosci*, *12*(6), 626-632. https://doi.org/10.1080/17470919.2016.1248787

Gordon, J., & Nyholm, S. (2021). Ethics of Artificial Intelligence. In *Internet Encyclopedia of Philosophy*

Greene, J. (2014). *Moral Tribes : Emotion, Reason and the Gap Between Us and Them*. Atlantic Books Ltd. http://kcl.eblib.com/patron/FullRecord.aspx?p=1486561

Greene, J. D. (2016). Our driverless dilemma. *Science*, *352*(6293), 1514-1515. https://doi.org/doi:10.1126/science.aaf9534

Grinbaum, A., Chatila, R., Devillers, L., Ganascia, J. G., Tessier, C., & Dauchet, M. (2017). Ethics in Robotics Research: CERNA Mission and Context. *IEEE Robotics & Automation Magazine*, *24*(3), 139-145. https://doi.org/10.1109/MRA.2016.2611586

Guardian, T. (2022). https://www.theguardian.com/business/2022/nov/03/bank-of-england-interest-rates-higher-uk-economy

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon/Random House.

Hansson, S. O., Belin, M.-Å., & Lundgren, B. (2021). Self-Driving Vehicles—an Ethical Overview. *Philosophy & Technology*, *34*(4), 1383-1408. https://doi.org/10.1007/s13347-021-00464-5

Horty, J. F. (2002). Skepticism and floating conclusions. *Artificial Intelligence*, *135*(1), 55-72. https://doi.org/https://doi.org/10.1016/S0004-3702(01)00160-6

IEEE. (2018). ETHICALLY ALIGNED DESIGN.
Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37-50. https://doi.org/https://doi.org/10.1016/j.bushor.2019.09.003
Kirkpatrick, K. (2015). The moral challenges of driverless cars. *Communications of the ACM*, *58*(8), 19-20. https://doi.org/10.1145/2788477
Lin, P. (2013). *The Ethics of Autonomous Cars*. The Atlantic. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/?utm_source=copy-link&utm_medium=social&utm_campaign=share
Lin, P., Lin, P., Abney, K., Bekey, G. A., & Ebscohost. (2012). *Robot ethics : the ethical and social implications of robotics*. MIT Press.
Lin, P., Lin, P., Jenkins, R. C., Abney, K., & Oxford University, P. (2017). *Robot ethics 2.0 : from autonomous cars to artifical intelligence*. Oxford University Press.
Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy & Technology*, *30*(4), 547-558. https://doi.org/10.1007/s13347-017-0284-0
March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, 102426. https://doi.org/https://doi.org/10.1016/j.joep.2021.102426
Millar, J. (2014a). *An ethical dilemma: When robot cars must kill, who should pick the victim?* Robohug.org. https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/
Millar, J. (2014b). *You Should Have a Say in Your Robot Car's Code of Ethics*. https://www.wired.com/2014/09/set-the-ethics-robot-car/#:~:text=You%20Should%20Have%20a%20Say%20in%20Your%20Robot,a%20robust%20standard%20of%20informed%20consent.%20U.S.%20DOT
Millard-Ball, A. (2018). Pedestrians, Autonomous Vehicles, and Cities. *Journal of Planning Education and Research*, *38*(1), 6-12. https://doi.org/10.1177/0739456x16675674
Nallur, V. (2020). Landscape of Machine Implemented Ethics. *Science and Engineering Ethics*, *26*(5), 2381-2399. https://doi.org/10.1007/s11948-020-00236-y
NHTSA. (2020). *Drunk Driving*. U.S. Department of Transportation. https://www.nhtsa.gov/risky-driving/drunk-driving
Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507. https://doi.org/https://doi.org/10.1111/phc3.12507
Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506. https://doi.org/https://doi.org/10.1111/phc3.12506
Nyholm, S., & Smids, J. (2020). Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*, *22*(4), 335-344. https://doi.org/10.1007/s10676-018-9445-9
OECD. (2019). *Artificial Intelligence in Society*. https://doi.org/https://doi.org/10.1787/eedfee77-en
OECD. (2023). *Artificial intelligence*. Organisation for Economic

Co-operation and Development (OECD). https://www.oecd.org/digital/artificial-intelligence/
Page, K. (2012). The four principles: can they be measured and do they predict ethical decision making? *BMC medical ethics*, *13*, 10-10. https://doi.org/10.1186/1472-6939-13-10
Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, *25*(3), 341-363. https://doi.org/10.1007/s10506-017-9210-0
Renieris, E. M., David Kiron, a., & Mills, S. (2022). *To Be a Responsible AI Leader, Focus on Being Responsible*. https://sloanreview.mit.edu/projects/to-be-a-responsible-ai-leader-focus-on-being-responsible/
Russell, S. (2015). Robotics: Ethics of artificial intelligence. *Nature*, *521*(7553), 415-418. https://doi.org/10.1038/521415a

Scharre, P. (2017). *The trouble with trying to ban 'killer robots'*. https://www.weforum.org/agenda/2017/09/should-machines-not-humans-make-life-and-death-decisions-in-war/

Schmidt, E. (2021). The AI Revolution and Strategic Competition with China. https://www.project-syndicate.org/commentary/ai-revolution-competition-with-china-democracy-vs-authoritarianism-by-eric-schmidt-2021-08

Schoder, D. (2018). *Does a tradeoff between inflation and unemployment exist?* American Economic Association. https://www.aeaweb.org/research/inflation-unemployment-retrospectives-milton-friedman-cruel-dilemma

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. https://doi.org/https://doi.org/10.1016/j.trc.2017.04.014

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in Machine Ethics: A Survey. *ACM Comput. Surv.*, *53*(6), Article 132. https://doi.org/10.1145/3419633

Tzachor, A., Whittlestone, J., Sundaram, L., & hÉigeartaigh, S. Ó. (2020). Artificial intelligence in a crisis needs ethics with urgency. *Nature Machine Intelligence*, *2*(7), 365-366. https://doi.org/10.1038/s42256-020-0195-0

UK, T. (2022). *Establishing a pro-innovation approach to regulating AI*. https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

The Universal Guidelines for Artificial Intelligence. https://thepublicvoice.org/ai-universal-guidelines/

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, *20*(1), 27-40. https://doi.org/10.1007/s10676-017-9440-6

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385-409. https://doi.org/10.1007/s11023-020-09537-4

Van Staveren, I. (2007). Beyond Utilitarianism and Deontology: Ethics in Economics. *Review of Political Economy*, *19*(1), 21-35. https://doi.org/10.1080/09538250601080776

Wallach, W., & Allen, C. (2009). *Moral machines : teaching robots right from wrong* (First O.U.P. paperback edition. ed.). Oxford University Press.

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Top Cogn Sci*, *2*(3), 454-485. https://doi.org/10.1111/j.1756-8765.2010.01095.x

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions* Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA. https://doi.org/10.1145/3306618.3314289

https://dl.acm.org/doi/pdf/10.1145/3306618.3314289

Yu, H., Z. Shen, C. Miao, C. Leung, Lesser, V. R., & Yang, Q. (2018). Building Ethics into Artificial Intelligence. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18),

Zhao, X., Malle, B. F., Li, J., Cho, M.-J., & Ju, W. (2016). *From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars* https://doi.org/10.4271/2016-01-0164

Zhu, A., Yang, S., Chen, Y., & Xing, C. (2022). A moral decision-making study of autonomous vehicles: Expertise predicts a preference for algorithms in dilemmas. *Personality and Individual Differences*, *186*, 111356. https://doi.org/https://doi.org/10.1016/j.paid.2021.111356